

Self-Supervised Learning of State Estimation for Manipulating Deformable Linear Objects

Mengyuan Yan¹, Yilin Zhu¹, Ning Jin², Jeannette Bohg¹

Abstract—We demonstrate model-based, visual robot manipulation of linear deformable objects. Our approach is based on a state-space representation of the physical system that the robot aims to control. This choice has multiple advantages, including the ease of incorporating physics priors in the dynamics model and perception model, and the ease of planning manipulation actions. In addition, physical states can naturally represent object instances of different appearances. Therefore, dynamics in the state space can be learned in one setting and directly used in other visually different settings. This is in contrast to dynamics learned in pixel space or latent space, where generalization to visual differences are not guaranteed. Challenges in taking the state-space approach are the estimation of the high-dimensional state of a deformable object from raw images, where annotations are very expensive on real data, and finding a dynamics model that is both accurate, generalizable, and efficient to compute. We are the first to demonstrate self-supervised training of rope state estimation on real images, without requiring expensive annotations. This is achieved by our novel self-supervising learning objective, which is generalizable across a wide range of visual appearances. With estimated rope states, we train a fast and differentiable neural network dynamics model that encodes the physics of mass-spring systems. Our method has a higher accuracy in predicting future states compared to models that do not involve explicit state estimation and do not use any physics prior, while only using 3% of training data. We also show that our approach achieves more efficient manipulation, both in simulation and on a real robot, when used within a model predictive controller.

I. INTRODUCTION

Manipulating deformable objects is an important but challenging task in robotics. It has a wide range of applications in manufacturing, domestic services and health care such as robotic surgery, assistive dressing, textile manufacturing or folding clothes [1], [2], [3]. Unlike rigid objects, deformable objects have a high-dimensional state space and their dynamics is complex and nonlinear. This makes state estimation challenging and forward prediction expensive.

We propose a vision-based system that allows a robot to autonomously manipulate a linear deformable object to match a visually provided goal state. Previous learning-based approaches towards this problem [4], [5] learn models in image space or a latent space, and do not incorporate any physics prior into the learning process. While conceptually

¹Mengyuan Yan, Yilin Zhu, and Jeannette Bohg are with School of Engineering, Stanford University. {mengyuan, ylzhu, bohg}@stanford.edu

²Ning Jin is with Calico Labs. This research is done during Ning's PhD at Stanford University. jennyjin@calicolabs.com

The Okawa Foundation and Toyota Research Institute ("TRI") provided funds to assist the authors with their research but this article solely reflects the opinions and conclusions of its authors and not of the Okawa Foundation, TRI or any other Toyota entity.

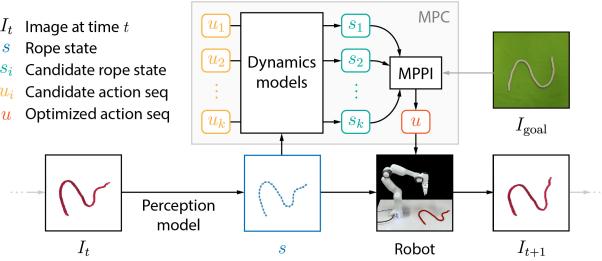


Fig. 1. Overview of our rope manipulation system. Given an image, our perception model estimates the explicit rope state. The state is refined by minimizing our proposed self-supervising objective w.r.t. the rope state. A dynamic model in rope state space predicts future states given the current state and hypothetical action sequences. MPPI is used to optimize action sequences according to the distance from predicted states to the goal state, also estimated from an image. The robot executes the first few actions and obtains a new observation. The state estimated from that image provides the input to MPPI to replan. This repeats until the goal state is reached.

these methods could be applied to other object classes, they suffer from low data efficiency and difficulty in generalization [6]. We take a different approach to this problem, based on an explicit state-space representation of the physical system. This choice has several advantages. First, it allows us to incorporate physics priors about the behaviour of a deformable object when it is manipulated, e.g. by reflecting a mass-spring system in the network structure. As we show in our experiments, such dynamics models produce more realistic predictions of the object's behaviour over a longer horizon than dynamics models learned directly from pixels. Second, explicit states are invariant to the appearance of the object and its environment. Therefore, dynamics learned in one setting can be directly used in other visually different settings. It also allows us to specify goal shapes with one rope that is then achieved with a rope of different length, thickness, and/or appearance. It is not obvious how to achieve this invariance with a method operating in a learned latent space or in pixel space. Finally, an explicit state-space representation more readily lends itself to manipulation planning and control especially when optimizing a sequence of actions. It is straight-forward to construct intuitive and informative losses for the optimization, as well as heuristics of promising action sequences to initialize the optimization.

The main challenge then becomes to estimate this explicit state from raw images. This task has previously been tackled by hand-engineered image processing algorithms, e.g. in [7]. Recently, Pumarola et al. [8] demonstrated explicit state estimation for deformable surfaces in simulation, where ground truth annotations are easily accessible. Such annotations are

expensive to obtain for real images. We overcome this problem by proposing a self-supervising learning objective that enable continuous, self-supervised training of deformable object state estimation on real images once the model is initialized with a small set of synthetic images.

We demonstrate the effectiveness of our method on the task of rope manipulation. We embed the learned perception model for explicit state estimation into a full system that includes a dynamic model in state space with physics priors, and *Model Predictive Control* (MPC), shown in Fig. 1. We quantitatively show that the proposed method is significantly more efficient in manipulating ropes to match specified goal configurations compared to previous methods that learn in pixel space.

Summarizing, the contributions of this work are:

- (i) We propose a novel self-supervising learning objective that enables training state estimation on real data, without requiring expensive ground truth annotations.
- (ii) We propose coarse-to-fine state estimation using hierarchical Spatial Transformer Networks (STN), which shows improved generalization compared to direct state estimation.
- (iii) We propose a novel dynamics model in state space that enforces physics priors for linear deformable objects. Our dynamics model has comparable performance to a physics simulation engine, while being significantly faster. Our model reduces prediction error by 68% while only using 3% of total training data, compared to a baseline dynamics model in pixel space [5].
- (iv) We demonstrate rope manipulation in both simulated and real environments. Quantitative comparisons in simulation show significantly more efficient manipulation of our method compared to a baseline.

II. RELATED WORK

a) Self-supervised State Estimation: While a lot of previous works learn dynamics models in image space or latent space using self-supervision, only a few have looked at self-supervised learning of explicit state estimation, such as object pose estimation. Wu et al. [9] used a differentiable renderer to convert predicted rigid object poses back to images, and compare with ground truth observations. Ehrhardt et al. [10] proposed two regularizing losses to enforce object trajectory continuity and spatial equivariance. Byravan et al. [11] train networks that learn robot’s latent link segmentation and pose space dynamics from point cloud time series, using a reconstruction loss for self-supervision. Our image-space loss is inspired by this line of works, but addresses much higher-dimensional linear deformable objects like ropes. We are able to train the perception network on real data with only self-supervision, provided that it has been warm started with a small set of rendered images.

b) Deformable Object Tracking: Several previous works have studied tracking of deformable objects given segmented point clouds [12], [13], [14]. On the high level, these tracking methods iteratively deform a pre-defined mesh model to fit the segmented point cloud, and use physical simulation to regularize the mesh deformation to have low

energy. There are two major limitations in the previous tracking methods: (1) the mesh configuration needs to be initialized manually or using algorithms engineered for specific cases. (2) Rope segmentation is required as input, usually achieved by color/depth filtering, assuming the foreground and/or background has a known solid color. Our state estimation method eliminates the need for manual initialization by introducing a perception neural network. We also alleviate the assumption of known foreground/background colors, only requiring that the foreground and background have good color contrast, to achieve self-supervision on real images. Our state estimation method is generalizable to a range of visually different objects and backgrounds.

c) Deformable Object Manipulation: There are two main branches of work in the area of deformable object manipulation other than rope manipulation. One branch of work focuses on clothing items [15], [16]. Researchers designed robot action strategies specific for each of the task stages, and formulated the object states as key point positions and object contours, to drastically simplify the complex problem. Our work tackles the simpler problem of linear object manipulation, proposing a much more generalizable solution, with task-independent state space and action space formulations. The other branch of work focuses on flexible beams or cables [17], [18]. These works assume that the elastic forces are much stronger than gravity or other external forces, such as friction, and the shape of the beam/cable can be fully determined by moving or twisting the two end-points. This model does not apply to soft ropes. The shape of ropes heavily depend on the manipulation history due to friction with the supporting plane. Furthermore, ropes are not fully controllable by just grasping the two end-points.

d) Rope Manipulation: Specific to rope manipulation, Yamakawa et al. [19] have demonstrated high-speed knotting, which depends on an accurate dynamic model of the robot fingers and the rope. Lee et al. [20] and Tang et al. [21] use spatial warping to transfer demonstrated manipulation skills to new but similar initial conditions. More related to our work, Li et al. [22] and Battaglia et al. [23] model ropes as mass-spring systems, and use graph networks to learn rope dynamics. However, they assume that the rope’s physical state is fully observable. Ebert et al. [5] learn a video prediction model, without any physical concept of objects or dynamics. However, a series of efforts [24], [25] has been made to find informative losses on images, which are required for long-horizon planning in the model predictive control framework. Wang et al. [26] embed images into a latent space associated with an action-agnostic transition model, plan state trajectories and servo the trajectory with an additional learned inverse model. While they achieve satisfactory result manipulating one particular rope, visually different ropes are not guaranteed to share the embedding space and transition model, making generalization difficult. Different from previous works, we directly estimate the ropes’ explicit states from images. This space lends itself more readily to efficient learning, flexible goal specification and manipulation planning than pixel space or latent spaces.

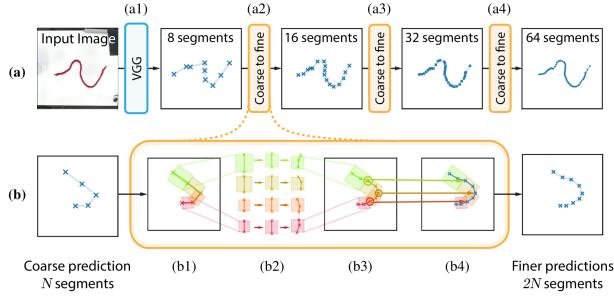


Fig. 2. Coarse-to-fine estimation of rope state. (a1) Given an input image, the neural network first estimates 8 straight segments. (a2-a4) The segment estimations are hierarchically refined using STNs. (b1) Square boxes are defined by the previously estimated segments. (b2) The boxes are used to extract regions from the VGG feature maps, and fed into a multi-layer perceptron to estimate the left endpoint, middle point, and right endpoint in each square region. The estimated points are concatenated(b3) and endpoints from neighboring regions are averaged(b4) to obtain a higher resolution estimate, based on twice the amount of segments that more closely model the shape of the rope.

III. METHOD

A flow chart of the full system at test time is shown in Fig. 1. We use a robot arm with gripper to manipulate a rope on the table. The task is to move the rope to match a desired goal state, specified by an image. At each time step, a *Convolutional Neural Network* (CNN) estimates the explicit rope state. The structure of this network is described in Sec. III-A. The network is first trained with rendered images, then finetuned on real images with our proposed self-supervising learning objective described in Sec. III-B. With the estimated states, we use *Model Predictive Path Integral Control* (MPPI) [27] in combination with a dynamics model to optimize a sequence of actions. Our proposed dynamic model is described in Sec. III-C. Details on the modified MPPI algorithm are described in Sec. III-D.

A. Coarse-to-fine State estimation

We formulate the problem of rope state estimation from an RGB image as estimating the positions of an ordered sequence of points on the rope. The rope state estimation problem has a divide-and-conquer structure, i.e., estimating the state of a segment of the rope is the same problem as estimating the state of the entire rope. To exploit this structure, we use STNs [28] to estimate the rope state in a coarse-to-fine manner as visualized in Fig. 2 (a). A VGG network [29] first estimates 8 straight segments that roughly approximate the shape of the rope. Using STNs, 8 square regions are extracted, one per segment. Within each extracted region, the network updates the position of the two endpoints and estimates the position of the middle point on the rope segment. The outputs are converted back to the original image coordinates, and endpoints from neighboring regions are averaged, so that the entire rope is now represented by 16 straight segments (see Fig. 2 (b)). New regions are extracted for each of the new segments, with higher spatial resolution. The process repeats until the rope is represented at sufficient resolution, in our case with 64 segments. The detailed network structure and parameters are described in

Appendix (I). Training and testing data for this and following networks are described in Appendix (V). Code will be made available upon publication.

B. Self-supervising learning objective

We train the neural network model with rendered spline curves as ropes, where ground truth rope states are easily available. However, real images look different from rendered images in many aspects, e.g. different lighting, distractor objects, occlusions through robot arms or, a different rope shape distribution. To close the reality gap without requiring ground truth annotations of rope states in real images, we propose a novel learning objective, consisting of a differentiable renderer and an image space loss to achieve self-supervision on real images (see Fig. 3).

Our method makes the assumption that the object has a good color contrast with the background, which is often the case for ropes or other linear deformable objects. Consider the simplifying case where the rope and the background each have a solid color. If we think of each pixel as a point in RGB space, all the pixels should form two clusters, one around the rope color and one around the background color. If we model the distribution in RGB space as a mixture of two Gaussians, and assign each pixel to the more probable Gaussian, the assignment variables will give us the segmentation mask of the rope versus background. When we estimate the rope configuration using the perception network, this estimate should agree with the color-based segmentation.

Clustering in RGB space can be achieved with the *Expectation-Maximization* (EM) algorithm for *Gaussian Mixture Models* (GMM). Given an initial estimate of GMM parameters Θ , i.e., the component weights α_k , means μ_k , and covariance matrices Σ_k , $1 \leq k \leq K$, the EM algorithm iterates between the E step, which updates the membership weights w_{ik} of data point x_i to cluster k , and the M step, which updates Θ . In the E step, membership weights are updated as:

$$w_{ik} = \frac{\alpha_k p_k(x_i | \mu_k, \Sigma_k)}{\sum_{m=1}^K \alpha_m p_m(x_i | \mu_m, \Sigma_m)} \quad (1)$$

where p_k and p_m are multivariate Gaussian densities. In the M step, GMM parameters are updated as:

$$\begin{aligned} \alpha_k^{new} &= \sum_{i=1}^N w_{ik}/N, \\ \mu_k^{new} &= \sum_{i=1}^N w_{ik} x_i / \sum_{i=1}^N w_{ik}, \\ \Sigma_k^{new} &= \sum_{i=1}^N w_{ik} (x_i - \mu_k^{new})(x_i - \mu_k^{new})^T / \sum_{i=1}^N w_{ik}. \end{aligned} \quad (2)$$

For rope state estimation, we model the distribution in RGB space with two Gaussians, thus $K = 2$. For each pixel with coordinate (u, v) , let $P(u, v)$ be its membership weight to the rope RGB cluster parameterized by μ_1 and Σ_1 , i.e., w_{i1} . Then $1 - P(u, v)$ is the membership weight of pixel (u, v) to the background RGB cluster parameterized by μ_2 and Σ_2 , i.e., w_{i2} . x_i refers to the RGB value of pixel

(u, v) . While the M step is straightforward to apply given $P(u, v)$, the per-pixel membership weights $P(u, v)$ should be expressed in terms of the estimated rope state, instead of i.i.d. per pixel. Thus, the E step does not apply as is.

We propose a differentiable renderer that links $P(u, v)$ to the rope state. The rope state is a sequence of 64 segments. We individually render each segment with end-points p_j , p_{j+1} to get $P_j(u, v)$, and take the pixel-wise maximum, $P(u, v) = \max_j P_j(u, v)$. Rendering of one segment is defined as

$$P_j(u, v) = \exp(-d_j(u, v)^2/\sigma^2) \quad (3)$$

where $d_j(u, v)$ is the distance of pixel (u, v) to its closest point on segment j . σ is a learnable parameter that controls the width of the rendered segments.

Given the initial state estimate from the neural network, we can compute $P(u, v)$ based on Eq. 3. The M step can be applied easily to compute the parameters of the Gaussian clusters in RGB space. We also follow the E step in Eq. 1 to calculate $P(u, v)^{new}$. Then, instead of directly using $P(u, v)^{new}$ for the next M step, we refine the estimated rope state by minimizing the distance between $P(u, v)$ and $P(u, v)^{new}$, defined as

$$\sum_{(u, v)} -\log [P(u, v)P(u, v)^{new} + (1 - P(u, v))(1 - P(u, v)^{new})]. \quad (4)$$

Thus, we adapt the EM algorithm for GMMs to minimize the above loss. Note that, the GMM parameters are only transient values estimated for each individual image. They are not memorized as parameters. Thus our method does not make any assumption about the distribution of rope colors or background colors, which was used in many previous works [12], [14], [15], [16]. Instead, we make the much weaker assumption that the rope has good color contrast with the nearby background.

To model occlusions, e.g. from the robot arm, we clip the gradient of this loss on each pixel $P(u, v)$ to be non-negative. In this way, we do not penalize pixels that belong to a rope segment according to the estimated rope state, but whose color belongs to the background color cluster, because that rope segment could be occluded.

The proposed loss can be used for either finetuning the perception network or for refining the rope state estimate at test time, without updating the network weights.

a) *Network finetuning with an automatic curriculum and temporal consistency:* While the self-supervising objective is generalizable across different visual appearances, it is not free of local minima. When the estimate from the perception network is not good enough, gradients of the proposed objective could lead the rope state into undesired local minima. When finetuning the perception network on real images, we want to prevent such undesired gradients from negatively affecting the network weights. We use an automatic curriculum based on the current loss of each training example. Training examples whose current loss is above a per-determined threshold are ignored during gradient updates. Since the selected examples are already very close

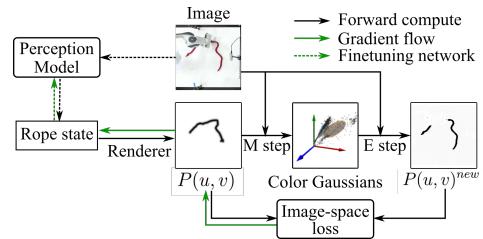


Fig. 3. Illustration of our proposed self-supervising learning objective. From an input image and initial rope state estimate, a differentiable renderer computes the membership weights $P(u, v)$. Then the M step produces GMM parameters, and the E step produces new membership weights $P(u, v)^{new}$. Given this, the image loss between $P(u, v)$ and $P(u, v)^{new}$ (Eq. 4) is minimized to either update the rope state, or to update the weights of the perception network (dashed green line).

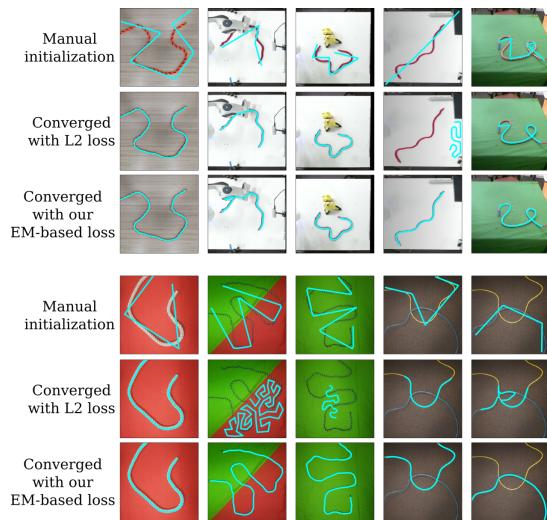


Fig. 4. Generalizability of self-supervising objectives in refining initial rope state estimations. Rope states are overlaid on the input images as blue lines. Our proposed objective is robust to lighting variation, robot occlusion, distracting objects, ropes/backgrounds with more than one color, and the presence of multiple ropes. Using the L2 loss failed to converge to the desired result on some real images.

to the true rope state thus having low loss, it is very unlikely their gradients will lead to wrong local minima. As the network learns, examples that originally have higher losses will improve, and their probability of falling into undesired local minima decreases. These examples will be included in the effective training set at a later point when their losses drop below the threshold.

In addition to using a curriculum, we also exploit temporal consistency in the recorded sequences to help the learning converge faster and better. If one frame has a loss below the threshold while its neighboring frame has a loss above the threshold, we take the predicted rope state from the better frame to guide the prediction on the worse frame. Exploiting temporal consistency in self-supervised training greatly improves the result, as shown in Fig 5. For more details about the finetuning algorithm, refer to Appendix (II).

b) *Generalization to more complex visual appearances:* Although we derived the objective for the simple case of rope and background each having a solid color, we note that this objective is also applicable if the rope or the

background is textured with several different colors, or when there are distractor objects or occlusion. We demonstrate a few examples in Fig. 4. To demonstrate the effectiveness of our objective independent of other components, we manually initialized the rope state estimation by clicking a few points on the images. During manipulation experiments, such initial estimates are provided by the perception network. The refined rope state estimate after convergence is shown in Fig. 4 (3rd and 6th row). We compare to the method generalized from [9], where the rendered grey scale image $P(u, v)$ is colored with the mean color of each cluster, and the L2 loss with the input image is used (Fig. 4 (2nd and 5th row)). Our proposed objective is robust to lighting variations, bi-colored ropes, bi-colored/textured backgrounds, distractor objects, multiple ropes, or occlusions. The baseline method does not always converge to the desired solution. The superior robustness of our method compared to using the L2 loss can be attributed to the different assumptions used by GMM clustering and K-means clustering. Using the same notation as above, the L2 loss can be written as:

$$\begin{aligned} & \sum_i \|x_i - w_{i1}\mu_1 - w_{i2}\mu_2\|^2 \\ = & \sum_i w_{i1}\|x_i - \mu_1\|^2 + w_{i2}\|x_i - \mu_2\|^2 + \sum_i w_{i1}w_{i2}\mu_1^T\mu_2. \end{aligned}$$

Since w_{i1}, w_{i2} are computed from Eq.3, $\sum_i w_{i1}w_{i2}$ only depends on the hyper parameter σ and the total length of the rope when approximated to the first order. Thus we do not consider the effects of the last term. The first two terms correspond to K-means clustering in the RGB space. K-means clustering has the following assumptions: (1) The variance of each cluster is roughly equal. (2) The number of data points in each cluster is roughly equal. (3) The distribution of each cluster is roughly isotropic (spherical). All of these assumptions can be broken in real images, e.g. when the rope or background has multiple colors, or when the variance of brightness of pixels is much larger than the variance of hue, due to lighting and shadows. On the other hand, our proposed loss is using GMM to cluster the RGB space, thus is more robust on real images.

C. A dynamics model with a physics prior

After rope states are estimated by the perception network and further refined with the proposed objective, a dynamics model is needed to predict future rope states given hypothetical actions, so that we can plan action sequences towards the goal.

Although physics-based simulators for deformable objects are available [30], we train a neural network dynamics model to speed up inference and support parallel processing. To encode the physics prior of linear objects, the neural network uses a bi-directional LSTM to model the structure of a chain-like mass-spring system. While LSTMs are usually used to propagate information in time, here we use it to propagate information along the rope's mass-spring chain in both directions. Recurrently applying the same LSTM cell to each node in the rope ensures that the same physical law

is applied, whether the node is closer to the endpoint or in the middle of the rope. Details of the network are described in Appendix (III). We also experimented with the recently proposed graph network [23] but found it less effective in propagating along a long chain of nodes. When generating training data, physical parameters used in the simulation are identified automatically using CEM on a small set of real data. Simulation sequences with random actions are generated and the model is trained on one-step prediction.

D. Rope manipulation with MPC

We use model predictive control to plan for a sequence of actions that takes the rope from the current configuration to the goal configuration. Both are estimated from input images. We formulate actions as first selecting a grasping point on the rope, and then selecting a 2D planar vector to move the gripper and the rope being grasped. This is different from the action space used in previous works [5], [4], where a grasping point is selected in image space, and a large portion of the action space will not make contact with the rope. Note that if our estimated rope configuration deviates from the real rope significantly, the robot may still fail to grasp the real rope. However, such cases rarely happen in our experiments, since minor errors from the perception network can be corrected in the refinement process with our proposed loss.

We adapt a sampling-based approach, MPPI [27], for planning actions to manipulate the rope. We perform a nested optimization to obtain the grasping points as well as movement trajectories. In the inner loop, we sample n movement trajectories of a given grasping point on the rope. These trajectories are rolled-out with our dynamics model over a time horizon T . The cost of each rope state along the trajectory is its distance to the goal state. The optimal trajectory per grasping points is computed as the cost-weighted average of the sampled trajectories, derived in [27]. For the outer loop, we sample grasping points on the rope and run the inner optimization loop for each in parallel. The grasping point with lowest cost of its optimal trajectory is selected.

Because the explicit rope states are available, defining an informative loss as well as sampling promising action candidates for MPPI is straight forward, compared to methods that operate in image space [5]. See Appendix (IV) for more details.

IV. EXPERIMENTS AND RESULTS

We evaluate each of the components described in the above section, and demonstrate that both our perception and dynamics model can be trained more effectively compared to baseline models that do not incorporate any prior structure. Our proposed self-supervising learning objective is able to transfer the perception model from simple rendered images to real images with unseen occlusions. In addition, the components work with each other to achieve efficient manipulation of ropes to match visually specified goals, both in simulation and on real robots.

TABLE I

ESTIMATION ACCURACY OF PERCEPTION NETWORKS. WE REPORT ROOT MEAN SQUARE OF THE EUCLIDEAN DISTANCE (IN METERS) BETWEEN ESTIMATED AND GROUND TRUTH POINT POSITIONS ON THE ROPE.

	Train	Test
Baseline: Direct Estimate	0.0104	0.0354
Ours: Coarse-to-Fine	0.0177	0.0231

A. Perception networks comparison

We evaluate estimation accuracy and generalization ability for two CNNs. The baseline model directly outputs 65 point coordinates from the last fully connected layer. We compare this to our proposed network that uses STNs for a coarse-to-fine estimation. Both models are trained on 10000 rendered images of b-spline curves. We report the training and evaluation loss for each network in Table I. Although the training loss for our network is larger than that of the baseline, our network achieves 30% less error on a held out test set, demonstrating better generalization due to our coarse-to-fine formulation.

Because the state space of a rope is very high dimensional, densely sampling in this space is difficult and would lead to a data set whose size is exponential in the number of rope points. Therefore, generalization as provided by the hierarchical STNs is very important for our problem.

B. Self-supervised finetuning

Since the robot arm is not modeled in the renderer, a network only trained on rendered images never sees the robot arm or the resulting occlusion of the rope, and thus it does not generalize well to real images (see Fig. 5). We use our proposed learning objective to finetune the parameters of our perception network on 5122 real images, without requiring annotations. We visualize the result after finetuning in Fig. 5. Ablation studies confirm that both automatic curriculum learning and temporal consistency brings significant improvements. The improvement is not sensitive to the selection of loss threshold, shown in Appendix (VI) B.

C. Learning dynamics models

We evaluate the long-horizon prediction accuracy of the learned dynamics model on both simulated and real data. We use two distance metrics for rope states: the average and the maximum deviation. Given a pair of rope states we first compute the Euclidean distance d_i for each pair of corresponding points, $i = 1, \dots, 65$. The average and maximum deviation are defined as $\text{mean}(d_i)$ and $\max(d_i)$. These metrics will be used for all the following experiments.

a) *Fast and accurate network:* We show the prediction accuracy of the neural network dynamics model on real data, and compare to the simulator it is trained from. As shown in Fig. 6 (left), the prediction accuracy of our neural network model is comparable to the simulation engine, with the initial state as well as ground truth states estimated from images. Noise in the lines are partially due to the noise of state estimation since the ropes are partially occluded.

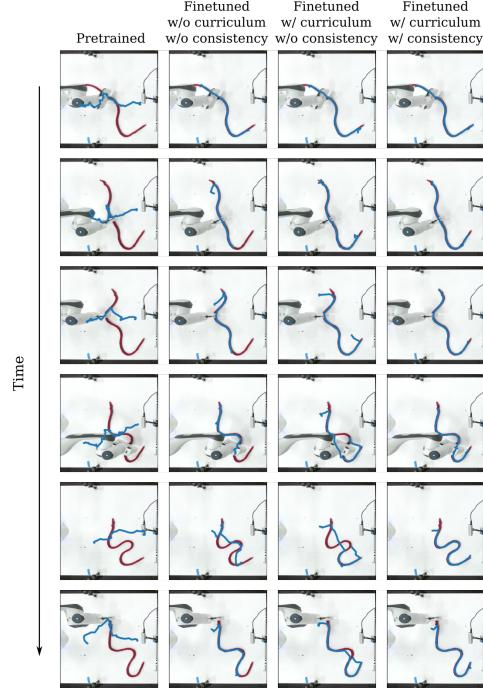


Fig. 5. Visualizations of finetuning the perception network weights with the proposed objective. First column: input images overlaid with state estimation before finetuning (in blue). Second column: model finetuned with the proposed objective, without curriculum learning or temporal consistency. Third column: model finetuned with curriculum learning but no temporal consistency. Fourth column: model finetuned with both curriculum and temporal consistency. The selected samples are from a training sequence.

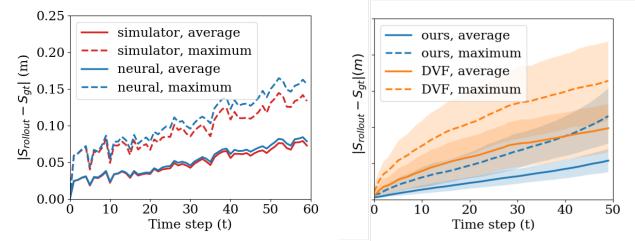


Fig. 6. Left: average (solid lines) and maximum (dashed lines) deviations from dynamic model predictions to estimated states from real observations, averaged over 27 sequences. Right (sharing y axis): average (solid lines) and maximum (dashed lines) deviations from dynamic model predictions to ground truth states in simulation. Lines represent the average from 200 sequences and shaded regions represent the standard deviation.

We expect the prediction accuracy to be further improved if also trained with multi-step prediction and finetuned on real data. The main advantage of the neural dynamics model is that it is significantly faster to predict, taking 0.03 second per action on average, compared to 1.15 second per action for the simulator. The neural network model is also readily parallelized on GPU with batch size up to 32000. Both aspects are beneficial to MPC, since a lot of mental rollouts are required in parallel.

b) *Benefit of incorporating a physics prior:* We further compare the long-horizon prediction accuracy of our neural network dynamics model with the visual dynamics model (DVF) from [5], on a batch of 200 sequences from the simulated dataset. Both models are trained with the same

dataset, except that DVF takes images, whereas our model takes explicit rope states. For each sequence, both models receive the same starting image and a sequence of 50 actions. The input state for our model is estimated from the image. For DVF, the model tracks 65 points on the rope by predicting heat maps in image space, and point positions are the expectations from predicted point distributions. As shown in Fig. 6 (right), our neural dynamics model is significantly more accurate than DVF. Note that both models are trained on a large dataset of 0.5M simulated actions, equivalent to at least 600 robot hours. Further experiments on 1-step prediction show that our proposed model can achieve 48% error reduction measured by maximum deviation, and 68% error reduction measured by average deviation, when only using 3% of the training data, compared to training the baseline model using the entire set (Appendix (VI)C). By trading off some level of generality of the method, we are able to incorporate the physics prior for deformable linear objects and achieve significant gain in data efficiency and generalization, manifested in the prediction accuracy and subsequent manipulation performance (Sec IV-D).

D. Manipulation results

We evaluate the performance of our system on the task of manipulating a rope on the table to match desired goal states. We compare our method with the baseline method [5], which uses MPC with the pixel distance cost. For the baseline, a task is specified by the start and goal positions of 11 equidistant points on the rope. To select promising grasping points from images, we compute the pixel-wise difference between the current observed image and the goal image, and only sample grasping positions where the two images are significantly different.

For quantitative evaluation, we run manipulation tasks in simulation, and select start states and goal states from randomly generated b-spline curves. Both our method and the baseline only see the rendered images. We report the distance to goal $L(t)$ as a function of time t , where $L(t)$ is the average deviation from the current rope state to the specified goal state. We show the mean and standard deviation over 100 independent experiments in Fig. 7. Our method achieves the goal state within 60 steps in most cases, and the remaining distance is very small, while the baseline method that operates in image space often cannot achieve the goal state within 100 steps, showing large residual distances at $t = 100$. For more visualizations refer to Appendix (V).

We also demonstrate rope manipulation on the real robot. We arrange the goal state of the rope to be an “S” shape, a “W” shape, or an “Ω” shape. Due to the very large state space of ropes, the network does not always generalize to the rope states during the manipulation tasks, which may be outside of the training distribution. Most of the overfitting is from the coarsest prediction layer, and the refinement layers generalize well if given a reasonable coarse prediction. This motivates us to further exploit temporal information during manipulation. Based on the latest estimated state s_t , the MPC plans the optimal action a_t . The learned dynamics

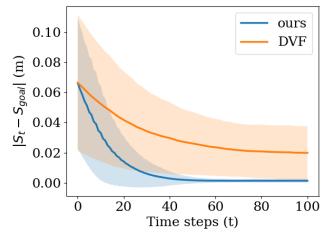


Fig. 7. Residual average deviation between the rope state at each time t and the goal state, for each manipulation method. Mean and standard deviation are obtained from 100 experiments of random start state and goal state pairs.

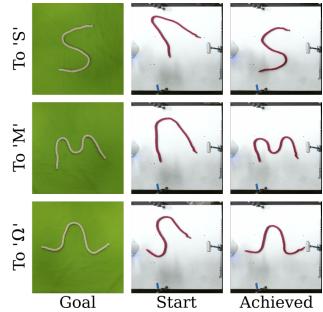


Fig. 8. From left to right: the start state, goal state, and achieved state after 40 steps. Goal images are rescaled for best comparison. Other images are taken by Kinect and projected to top-down view.

model predicts the next state \hat{s}_{t+1} . \hat{s}_{t+1} is subsampled into 8 segments to feed into the CNN’s first STN, together with the next image I_{t+1} , and the CNN refines \hat{s}_{t+1} into the next estimated state s_{t+1} . Using this temporal information greatly improved generalization of our perception method. Appendix (VI)A shows the tracking results where an intersection is made with the rope, which is never seen during training.

To highlight one benefit of using explicit state representations, we use a different rope on a different background to demonstrate goals. The two ropes have different appearance as well as different lengths and thicknesses. The L2 loss between goal and observed images would not be informative for video prediction methods, and it would be hard to embed the goal image into the latent space of the manipulated rope, if using [26]. Our method successfully finished all 3 tasks, visualized in Fig. 8. Also see supplementary material for robot manipulation videos.

V. CONCLUSION

We demonstrated model-based, visual robot manipulation of deformable linear objects. Our forward model makes explicit estimation of rope states from images, and learns a dynamic model in state space. We proposed a self-supervising learning objective to enable self-supervised continuous training of rope state estimation on real data, without requiring expensive annotations. Our objective is generalizable across a wide range of visual appearances. We also proposed coarse-to-fine state estimation using hierarchical STNs that greatly improves generalization to unseen rope shapes. With access to the rope’s explicit state, we are able to incorporate physics priors, e.g., the structure of mass-spring systems, into the

design of the network structure for dynamics models, in addition to using physical simulation for data generation. We demonstrated that our dynamics model using bi-directional LSTM has reduced prediction errors by up to 68% while only using 3% of total training data, compared to a baseline dynamics model in pixel space, and that our method achieves more efficient manipulation in matching visually specified goals. Although we only demonstrated manipulation on a horizontal plane, our method can be extended to 3D manipulation tasks with minor modifications. In future works, instead of assuming a known table height, we can use the estimated 2D states to extract depth values for the Kinect depth images to reproject the estimated points to 3D.

For future directions, it would be interesting to explore using our self-supervising learning objective to continuously train the perception and dynamics network while the robot is performing manipulation tasks, similar to DAGGER [31]. Our neural network dynamics model can be extended to have object's physical properties as latent variables, such that the dynamics model can adapt quickly to ropes/wires with different physical properties, by updating the latent variables instead of network weights. Finally, we would also like to extend this method to deformable objects with even higher dimensional state spaces, such as clothing. Possible directions include adapting the proposed perception method to estimate the contour of clothes, and extending the proposed bi-directional LSTM to a 2-dimensional bi-directional LSTM to model the dynamics of clothes.

ACKNOWLEDGMENTS

We thank Professor Ronald Fedkiw for useful discussions on deformable object simulation.

REFERENCES

- [1] S. Leonard, A. Shademan, Y. Kim, A. Krieger, and P. C. Kim, "Smart tissue anastomosis robot (star): Accuracy evaluation for supervisory suturing using near-infrared fluorescent markers," *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1889–1894, May 2014.
- [2] A. Clegg, "Learning to dress: synthesizing human dressing motion via deep reinforcement learning," *ACM Trans. Graph.*, vol. 37, pp. 179:1–179:10, 2018.
- [3] Y. Li, Y. Yue, D. Xu, E. Grinspun, and P. K. Allen, "Folding deformable objects using predictive simulation and trajectory optimization," *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 6000–6006, Sep. 2015.
- [4] A. Nair, D. Chen, P. Agrawal, P. Isola, P. Abbeel, J. Malik, and S. Levine, "Combining self-supervised learning and imitation for vision-based rope manipulation," *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2146–2153, May 2017.
- [5] F. Ebert, C. Finn, S. Dasari, A. Xie, A. X. Lee, and S. Levine, "Visual foresight: Model-based deep reinforcement learning for vision-based robotic control," *arXiv: 1812.00568*, 2018.
- [6] A. Kloss, S. Schaal, and J. Bohg, "Combining learned and analytical models for predicting action effects," *arXiv:1710.04102*, 2017.
- [7] T. Morita, J. Takarnatsu, K. Ogawarat, H. Kiniuratt, and K. Ikeuchi, "Knot planning from observation," *IEEE International Conference on Robotics and Automation (ICRA)*, vol. 3, pp. 3887–3892, May 2003.
- [8] A. Pumarola, A. Agudo, L. Porzi, A. Sanfeliu, V. Lepetit, and F. Moreno-Noguer, "Geometry-aware network for non-rigid shape prediction from a single view," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [9] J. Wu, E. Lu, P. Kohli, B. Freeman, and J. Tenenbaum, "Learning to see physics via visual de-animation," *International Conference on Neural Information Processing Systems (NIPS)*, pp. 153–164, Dec. 2017.
- [10] S. Ehrhardt, A. Monszpart, N. J. Mitra, and A. Vedaldi, "Unsupervised intuitive physics from visual observations," *Asian Conference on Computer Vision*, pp. 700–716, 2018.
- [11] A. Byravan, F. Leeb, F. Meier, and D. Fox, "SE3-pose-nets: Structured deep dynamics models for visuomotor control," *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1–8, May 2018.
- [12] J. Schulman, A. Lee, J. Ho, and P. Abbeel, "Tracking deformable objects with point clouds," *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1130–1137, May 2013.
- [13] A. Petit, V. Lippiello, and B. Siciliano, "Real-time tracking of 3d elastic objects with an rgb-d sensor," *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3914–3921, Sep. 2015.
- [14] T. Tang, Y. Fan, H. Lin, and M. Tomizuka, "State estimation for deformable objects by point registration and dynamic simulation," *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2427–2433, Sep. 2017.
- [15] A. Doumanoglou, J. Stria, G. Peleka, I. Mariolis, V. Petrik, A. Karagakos, L. Wagner, V. Hlaváč, T. Kim, and S. Malassiotis, "Folding clothes autonomously: A complete pipeline," *IEEE Transactions on Robotics*, vol. 32, pp. 1461–1478, Dec 2016.
- [16] Y. Li, Y. Wang, Y. Yue, D. Xu, M. Case, S.-F. Chang, E. Grinspun, and P. K. Allen, "Model-driven feedforward prediction for manipulation of deformable objects," *IEEE Transactions on Automation Science and Engineering*, vol. 15, pp. 1621–1638, 2016.
- [17] M. Moll and L. E. Kavraki, "Path planning for deformable linear objects," *IEEE Transactions on Robotics*, vol. 22, pp. 625–636, 2006.
- [18] O. Roussel, A. Borum, M. Taïx, and T. Bretl, "Manipulation planning with contacts for an extensible elastic rod by sampling on the submanifold of static equilibrium configurations," *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3116–3121, May 2015.
- [19] Y. Yamakawa, A. Namiki, and M. Ishikawa, "Dynamic high-speed knotting of a rope by a manipulator," *International Journal of Advanced Robotic Systems*, vol. 10, p. 361, 2013.
- [20] A. X. Lee, S. H. Huang, D. Hadfield-Menell, E. Tzeng, and P. Abbeel, "Unifying scene registration and trajectory optimization for learning from demonstrations with application to manipulation of deformable objects," *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4402–4407, Sep. 2014.
- [21] T. Tang, C. Liu, W. Chen, and M. Tomizuka, "Robotic manipulation of deformable objects by tangent space mapping and non-rigid registration," *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2689–2696, Sep. 2016.
- [22] Y. Li, J. Wu, J.-Y. Zhu, J. B. Tenenbaum, A. Torralba, and R. Tedrake, "Propagation networks for model-based control under partial observation," *IEEE International Conference on Robotics and Automation (ICRA)*, May 2019.
- [23] P. Battaglia, R. Pascanu, M. Lai, D. J. Rezende, and K. kavukcuoglu, "Interaction networks for learning about objects, relations and physics," *International Conference on Neural Information Processing Systems (NIPS)*, Dec. 2016.
- [24] F. Ebert, C. Finn, A. X. Lee, and S. Levine, "Self-supervised visual planning with temporal skip connections," *Conference on Robot Learning*, pp. 344–356, 2017.
- [25] F. Ebert, S. Dasari, A. X. Lee, S. Levine, and C. Finn, "Robustness via retrying: Closed-loop robotic manipulation with self-supervised learning," *Conference on Robot Learning*, vol. 87, pp. 983–993, 2018.
- [26] A. Wang, T. Kurutach, K. Liu, P. Abbeel, and A. Tamar, "Learning robotic manipulation through visual planning and acting," *Proceedings of Robotics: Science and Systems*, 2019.
- [27] G. Williams, P. Drews, B. Goldfain, J. M. Rehg, and E. A. Theodorou, "Aggressive driving with model predictive path integral control," *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1433–1440, May 2016.
- [28] M. Jaderberg, K. Simonyan, A. Zisserman, et al., "Spatial transformer networks," *International Conference on Neural Information Processing Systems (NIPS)*, pp. 2017–2025, Dec. 2015.
- [29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *International Conference on Representation Learning (ICRL)*, 2015.
- [30] Physbam: physically based animation. <http://physbam.stanford.edu/>.
- [31] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," *International Conference on Artificial Intelligence and Statistics*, pp. 627–635, 2011.

Self-Supervised Learning of State Estimation for Manipulating Deformable Linear Objects

Appendix

Mengyuan Yan, Yilin Zhu, Ning Jin, Jeannette Bohg

I. PERCEPTION MODEL

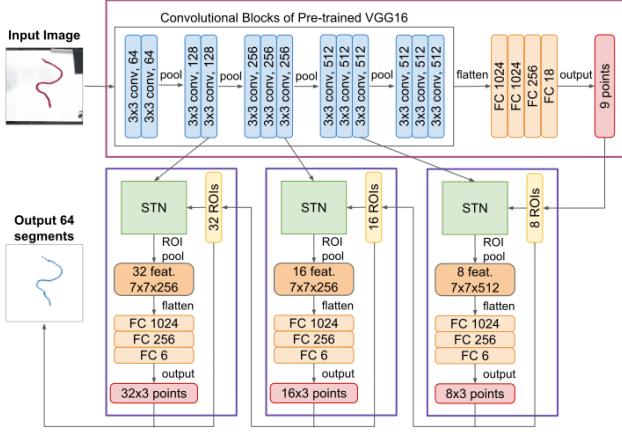


Fig. 1. Detailed architecture of our perception network. Top: Initial coarse prediction of 8 segments on the rope. We first run input images through the convolutional blocks of a pretrained VGG16 network to extract feature maps at different resolutions, and the feature map from the last convolutional layer then goes through 4 fully connected layers to output initial estimation for the 9 points that define the 8 ROIs. Bottom: Hierarchical refinement of the estimation via spatial transformer networks. Starting from the initial coarse prediction from above, we iteratively refine the estimation, where each time we double the number of ROIs. We run a total of 3 steps, and end up with 64 segments as our final output estimation.

In this section, we describe the details of our perception model, which can be broken down into two components: 1) an initial coarse estimation component using a pretrained VGG16 [?] network and a few fully connected layers (Fig. 1 top) and 2) a hierarchical refinement component using spatial transformer networks on feature maps from the VGG network (Fig. 1 bottom).

As shown in Fig. 1 top, the initial coarse estimation takes in 224×224 images and outputs 9 points (18 numbers in 2D) that define 8 segments to approximate the shape of the rope. We first run the image through the convolutional blocks of a pretrained VGG network to extract feature maps at different resolutions. Then the feature map from the last convolutional layer goes through 4 fully connected layers of size 1024, 1024, 256, and 18 to output prediction for the 9 points. During training, the VGG network is frozen and only the fully connected layers are updated.

The initial estimation is then refined hierarchically as shown in Fig. 1 bottom. Starting from the initial 8 segments,

we repeatedly refine the estimation by doubling the number of ROIs each time, i.e., $8 \rightarrow 16 \rightarrow 32 \rightarrow 64$. At step i ($i = 1, 2, \dots$), we start with $8 \times 2^{i-1}$ ROIs, and for each ROI, we crop out feature maps from the last convolutional layer in the $(5 - i)$ -th convolutional block of the pretrained VGG network, and transform the cropped feature maps into 7×7 spatial resolution using the STN. The features are then flattened and go through 3 more fully connected layers of size 1024, 256, and 6 to predict $(8 \times 2^{i-1}) \times 3$ points, where each triplet of points is predicted independently from the corresponding feature, and refines the corresponding segment by estimating its start-point, mid-point, and end-point. These estimations are then combined, where end-points of previous ROI and start-points of the next ROIs are averaged, and mid-points are kept as estimated. After step i , we get 8×2^i ROIs for the next step. We run a total of 3 steps, and end up with 64 segments as our final output for the rope state estimation.

II. SELF-SUPERVISED NETWORK FINETUNING

The details of the finetuning algorithm is summarized in Algorithm 1. We optimize the network weights with Adam optimizer to minimize the self-supervising learning objective (image loss). To incorporate automatic curriculum learning, we set a loss threshold, which is always a negative number. All training samples that have a smaller loss than the threshold are taken into account for the gradient updates, and other samples are omitted. Experimentally, we set the loss threshold to include about top 10% training samples in the first epoch. If the number of training samples below the threshold stops growing for 2 epochs, we increase the threshold by multiplying 0.98.

To incorporate temporal consistency, we restructure the dataset so that it consists of pairs of neighboring RGB images. During training, for each pair of images, if one of them has image loss below the threshold and the other has image loss above the threshold, the worse one is trained to minimize the L2 loss with the prediction of the better one as its "ground truth label", and the better one is trained to minimize the image loss. If both images have image loss below the threshold, then both images are trained with image loss. If both images have image loss above the threshold, they are omitted for this epoch.

III. NEURAL NETWORK DYNAMICS MODEL

In this section we give the equations that describe our bi-directional LSTM model. For each of the 65 nodes on the

Algorithm 1 Algorithm for finetuning perception network.

Dataset consists of pairs of neighboring RGB images.
hyper-parameters: $loss_thresh$, batch size=48

```

for epoch = 1, ..., N do
    Shuffle dataset
    if num_trained did not increase then
        loss_thresh ← 0.98 loss_thresh
    end if
    num_trained ← 0
    for each batch do
         $I_i, i = 0, \dots, 47$  are RGB images.
         $I_{2j}$  and  $I_{2j+1}$  are neighboring images.
        Forward pass  $S_i = \text{Net}(I_i)$ 
        Evaluate loss  $L_i = \text{ImageLoss}(S_i, I_i)$ 
        Total training loss  $L \leftarrow 0$ 
        for  $i = 0, \dots, 47$  do
            if  $L_i < loss\_thresh$  then
                 $L = L + L_i$ 
                num_train = num_train + 1
            end if
        end for
        for  $j = 0, \dots, 23$  do
            if  $L_{2j} < loss\_thresh$  and  $L_{2j+1} > loss\_thresh$  then
                 $L = L + l2(\text{StopGradient}(S_{2j}) - S_{2j+1})$ 
            end if
            if  $L_{2j} > loss\_thresh$  and  $L_{2j+1} < loss\_thresh$  then
                 $L = L + l2(\text{StopGradient}(S_{2j+1}) - S_{2j})$ 
            end if
        end for
        Update network with gradients of  $L$ .
    end for
end for

```

rope, its input contains its position $p_i \in \mathbb{R}^2$, action $a_i \in \mathbb{R}^2$ (which is 0 unless the action applies on this node), and an indicator $f_i = \mathbb{1}(|a_i| > 0)$. The inputs are concatenated into $x_i = (p_i, a_i, f_i) \in \mathbb{R}^5$ for $i = 1, \dots, 65$. p_i are retrieved from simulation during training and estimated from images during evaluation.

The bi-directional LSTM is constructed as

$$\begin{aligned}
h_1^L &= 0, \\
z_i^L, h_{i+1}^L &= \text{LSTM}(x_i, h_i^L), i = 1, \dots, 65, \\
h_{65}^R &= 0, \\
z_i^R, h_{i-1}^R &= \text{LSTM}(x_i, h_i^R), i = 65, \dots, 1, \\
y_i &= w_L z_i^L + w_R z_i^R + w_I x_i,
\end{aligned}$$

where the superscript L denotes the LSTM propagating from node 1 to node 65, and the superscript R denotes the LSTM in the reverse direction. h_i^L and h_i^R are the memory units, z_i^L and z_i^R are the output units. The LSTM cell has one layer with 256 units and ReLu6 activation. LSTM outputs z_i^R and z_i^L are concatenated together with the input x_i , and fed into one more linear layer to predict the position p_i^{out} of the i th node after the action is performed. The LSTM can be applied repeatedly on a sequence of actions for long-horizon prediction.

IV. HEURISTICS USED FOR ACTION PLANNING

In this section we describe the heuristics used to generate candidate action sequences for MPPI in our manipulation

experiments. These heuristics are made possible because we explicitly estimate rope states for the current observation and the goal image.

At each time step t , with current rope state $s_t = \{p_{i,t}, i = 1, \dots, 65\}$, and goal state $s_{\text{goal}} = \{p_{i,\text{goal}}, i = 1, \dots, 65\}$, we densely sample every other node, i.e. node 1, 3, 5, ..., 65 as candidate grasping points. Then, for each candidate grasping point, we generate 30 sequences each containing 10 displacement vectors. To generate each sequence starting at point i , we first calculate the unit vector e_1 in the direction of $p_{i,\text{goal}} - p_{i,t}$, and e_2 such that $e_2 \perp e_1$. all actions are initialized as $0.8a_{\max}e_1$, where a_{\max} is the maximum magnitude of displacement vectors. In case $d_{i,t} = |p_{i,\text{goal}} - p_{i,t}| < 8a_{\max}$, i.e., the goal position can be reached before the 10 action finishes, the final actions are clipped in magnitude so that the sequence should bring the grasping point to $p_{i,\text{goal}}$. Then we add exploration noises to the sequence. Exploration noise is specified by three random variables δ_x , δ_y and δ_c . Intuitively, δ_x and δ_y moves the endpoint of the 10-action trajectory in the 2D plane, while δ_c modifies the trajectory from a straight line to a curve, without changing the endpoint. Mathematically, the 10 displacement vectors $a_{i,1}, a_{i,2}, \dots, a_{i,10}$ are modified as

$$a_{i,t} = a_{i,t}^{\text{init}} + (\delta_x, \delta_y) + \delta_c \cos(t\pi/10)e_2. \quad (1)$$

V. DATA COLLECTION

a) Rendered images for perception model pretraining: We generated a dataset of 10000 rendered images, each containing a randomly generated b-spline curve with six control points. The b-spline curve is rendered with solid red color on white background. 65 equidistant points are extracted from the spline as ground truth annotation of rope states. The length of the generated ropes range from 0.63m to 1.25m.

b) Simulated dataset for training dynamics models: We generated a dataset of simulated rope manipulation sequences with ground truth rope states, actions and rendered images. The start state of each sequence is sampled from the dataset of 10000 b-splines used above. For each start state, we generate a 100-step manipulation sequence. At each time step, a virtual robot grasps the rope at a random point and moves that point with a randomly generated displacement vector. The magnitude of random displacements is between 1 and 3cm. After the rope configuration is computed by the simulator, we render the rope from a top-down view using POV-Ray [?]. The images are only used for training the baseline video prediction model [?]. A total of 5800 sequences are generated. We use 5112 sequences for training the dynamics neural network and the rest for evaluations.

c) Real dataset for perception model finetuning: We also collected a dataset of real rope manipulation sequences in a similar manner. We used a Franka Panda robot arm [?] with a parallel gripper to execute actions, and a Kinect camera to collect RGB images. The Kinect camera coordinate frame is calibrated to the robot base frame. Images collected from the Kinect camera are projected to top-down views

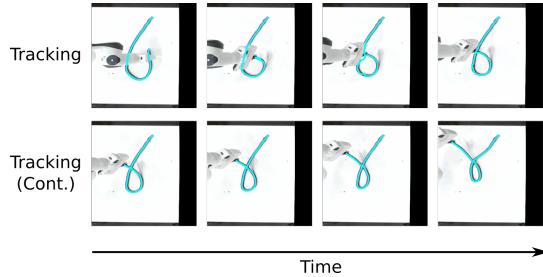


Fig. 2. State estimation and tracking results on a manipulation sequence, where the robot actions are designed by the authors to make an intersection with the rope.

using this transformation. No depth data is required for the calibration or projection. At the beginning of every recorded sequence, a human operator arranges the rope on the table. At each time step, the robot can choose to pick up the rope, move the gripper with rope being grasped, or release the rope and retract the arm from the table. For selecting the grasp location, a color filter is used on the Kinect image to segment the rope, and a random pixel in the rope segment is selected. Displacement vectors for moving the gripper are selected at random as long as the gripper stays within the robot’s workspace. Images are taken after each actions. For ease of transferring the pretrained network, the robot can move the gripper at most 5 times before releasing and retracting, so that we have un-occluded images of the rope spread through the recorded sequences. The real rope has length 1m, and actions range between 1 and 3cm. We collected a total of 27 sequences, with 26 sequences totaling 5118 images for training, transferring the pretrained perception model to the real environment using the self-supervising objective, and 1 sequence with 201 images for validation of the perception model. This dataset is also used to test the long-horizon prediction accuracy of our dynamics model, shown in Fig.6 (left) of the main paper. When using this data to evaluate the dynamics model, we use the state estimated by the finetuned perception network, further refined by directly optimizing the image loss w.r.t states, as the ground truth states.

VI. ADDITIONAL EXPERIMENT RESULTS

A. Tracking a rope making intersection

Visualization of the state estimation and tracking result on one manipulation sequence is shown in Fig. 2. The robot actions are designed by the authors to make an intersection (topological change). The tracking method is described in Sec. IV(D) of the paper. This result demonstrates that our perception method can also be used in more complex rope manipulation tasks, such as knotting.

B. Network finetuning with curriculum learning

We provide more detailed experiment results on the effects of choosing different error thresholds for the automatic curriculum learning. We choose error thresholds that correspond to the top 5%, 10%, 20%, and 40% quantile image loss among the training dataset before finetuning. The final average image losses after convergence are shown in Fig. 3.

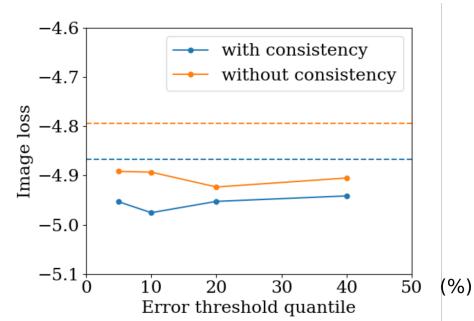


Fig. 3. Final image losses on the training set for finetuning with different error thresholds in automatic curriculum learning. The dashed straight lines are the final image losses when trained without using curriculum learning.

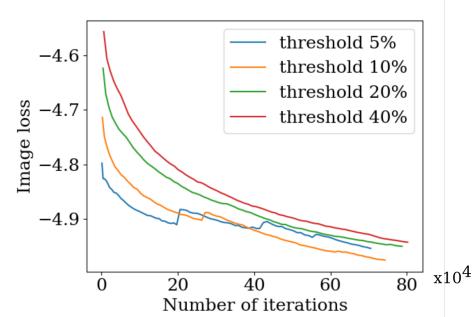


Fig. 4. The training loss curves for finetuning the network with both automatic curriculum learning and temporal consistency. The reported image losses are averaged over the effective training set, i.e. samples whose loss is below the error threshold. Stairs in the training loss curves correspond to when the error threshold is automatically adjusted to include more training samples.

The marked dots represent training results for different error thresholds, and the dashed lines provide baseline results where curriculum learning is not used. The figure shows it is best to use an error threshold between the 10% to 20% quantile of the training dataset, however the differences among different error thresholds are small compared to the improvement upon the baselines, thus demonstrating that using automatic curriculum learning is effective for a wide range of error thresholds used.

We also include the training curves when training with both automatic curriculum learning and temporal consistency, where the error threshold for curriculum learning varies. The training curves are shown in Fig. 4. The reported image losses are averaged over effective training samples, i.e. samples whose losses are below the error threshold. Therefore, the initial training losses for stricter error thresholds are lower. When training with a stricter error threshold, there are stairs in the training curve, corresponding to when the numbers of effective training samples have stopped increasing, and the error thresholds are automatically updated to include more training samples.

C. Data efficiency of dynamics models

We provide more experiment results to demonstrate that our dynamics model using bi-directional LSTM has much

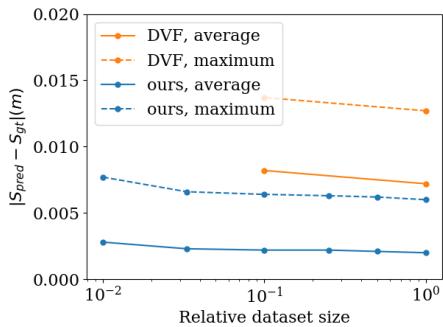


Fig. 5. Average and maximum deviation from predicted states to the ground truth on the evaluation set, when the training data size varies for both our LSTM model and the baseline model. The data sizes are measured as relative sizes compared to the largest dataset we generated, which contains 0.5M simulated actions.

higher data efficiency compared to the baseline model [?]. In Fig. 5, we plot the average and maximum deviation from the predicted states to the ground truth states on the evaluation set, when the training data size varies. The training data sizes are reported as relative sizes compared to the biggest dataset we generated, which contains 0.5M simulated actions, as described in the previous section. The results show that our dynamics model is able to train well with only 3% of the total training data, only showing overfitting and worsened results when training data size reduces to 1%. On the other side, the baseline model would need larger datasets to continue to improve its performance, which we do not have enough resources to generate.

D. Simulated manipulation experiments visualization

We include visualizations of the simulated manipulation experiments, described in Sec. IV(D) of the main paper. The start state, specified goal state, and states achieved by each method after 100 actions are shown in Fig 6. Our method achieves states much closer to the specified goals.

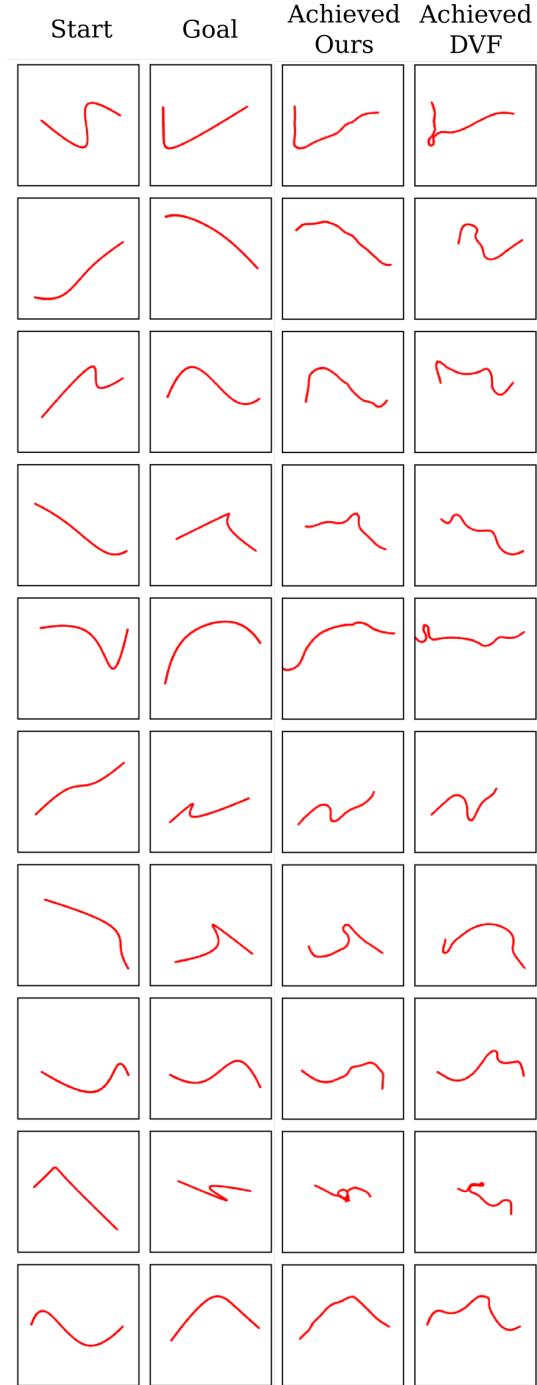


Fig. 6. 10 example manipulation experiments in simulation. From left to right: the start state, the specified goal state, state achieved with our method at $t = 100$, and state achieved with the baseline method(DVF)[?] at $t = 100$.