All You Need is LUV: Unsupervised Collection of Labeled Images Using UV-Fluorescent Markings

Brijen Thananjeyan*, Justin Kerr*, Huang Huang, Joseph E. Gonzalez, Ken Goldberg

***Abstract*— Learning-based perception systems in robotics often requires large-scale image segmentation annotation. Current approaches rely on human labelers, which can be expensive, or simulation data, which can visually differ from real data. This paper proposes Labels from UltraViolet (LUV), a novel framework that enables rapid, automated, inexpensive, high quality data collection in real. LUV uses transparent, UV-fluorescent paint with programmable UV LEDs to collect paired images of a scene in standard and UV lighting. This makes it possible to autonomously extract segmentation masks and keypoints via color thresholding. We apply LUV to a suite of diverse robot perception tasks: locating fabric keypoints, cable segmentation, and surgical needle detection to evaluate its labeling quality, flexibility, and data collection rate. Results suggest that LUV is 180-2500 times faster than a human labeler across the tasks while retaining accuracy and strong task performance. Code, datasets, visualizations, and supplementary material can be found at `https://sites.google.com/berkeley.edu/luv`.**

I. Introduction

Supervised learning of image segmentation is a popular technique for training perception and planning systems for robots, with encouraging results in applications such as autonomous driving [13, 21, 34], robot object grasping [5, 11, 17, 21, 26], deformable manipulation [12, 15, 22, 36, 41, 47, 48], and robot-assisted surgery [29, 45]. Supervised learning requires labeled data, and a common approach is for humans to hand-label images with segmentation masks, keypoints, and class labels [15, 19, 45]. However this is time-consuming, error-prone, and expensive [34], especially when dense annotations are required [5, 11, 12, 21, 41]. An alternative approach is to use simulated data, where data annotation can be densely and autonomously generated at scale at relatively low cost [5, 12, 17, 18, 21, 26, 41].

In this paper, we present Labels from UltraViolet (LUV) (Figure 1), a novel framework for rapidly and automatically collecting inexpensive and high quality ground-truth annotations without human labels. LUV uses an array of ultraviolet lights placed around a manipulation workspace that can be switched automatically. We mark objects or keypoints in the scene with transparent, ultraviolet fluorescent paints that are nearly invisible in visible light but highly visible under ultraviolet radiation. For physical configurations, LUV takes two images: one with standard lighting and one with the ultraviolet lights turned on. LUV provides precise labels for the standard image by using the paired ultraviolet image and trains a network on the resulting dataset to make predictions on subsequent scenes without UV paint under standard lighting. Since submitting this paper, we learned of impressive prior work by Takahashi and Yonekura [43]. They demonstrated compelling image segmentation results using UV-fluorescent markers to segment fluid, powders, and cloth keypoints by developing a hand-held device which strobes UV lights to collect real-time image annotations. This paper extends the concept to make it more easily accessible for self-supervised data collection. LUV has several desirable qualities:

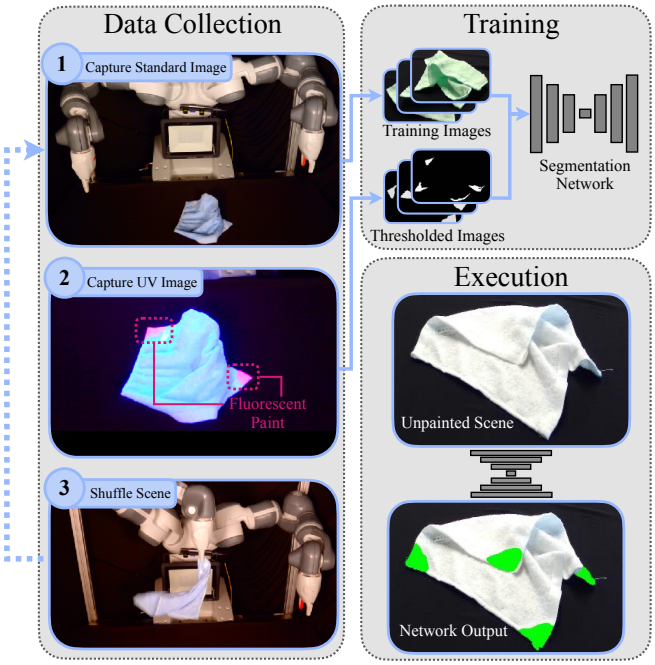

Fig. 1: **Framework overview. Data collection:** LUV collects paired images in standard and UV lighting. Relevant keypoints (in this case, corners) are coated with transparent, UV-fluorescent markings which are used to extract annotations from the UV images for the standard images. **Training:** The annotations are used to train a segmentation network to predict masks from images under standard lighting. In contrast to prior approaches to obtain segmentation labels for images, LUV requires no human annotator or simulator. **Execution:** During execution, the trained network is takes in images of unpainted objects and returns corner locations.

1) **accurate** segmentation masks and keypoints on *real images*,
2) **flexibility** to a wide variety of materials and tasks,
3) **rapid** data collection with *no human annotation,*
4) **inexpensive** setup with off-the-shelf parts costing less than $300 total.

To quantitatively evaluate these properties, we apply LUV to 3 real-world perception tasks in robot manipulation: locating fabric keypoints [12, 21, 36, 37], cable segmen-

* Equal contribution
The AUTOLab at UC Berkeley (automation@berkeley.edu)

tation [41], and surgical needle segmentation [38, 42, 45]. For each, we report the speed of data collection, qualitative invisibility of markings, transferability to unpainted images, and correspondence to human ground truth labels. Because we are able to collect data much more efficiently than in prior work, we study more visually complex variants of these problems than previously considered.

This paper makes the following contributions:

1) LUV, an easy-to-setup, inexpensive framework for rapid, automated and high quality ground truth image annotations collection that is 180-2500 times faster than a human labeler.
2) A user-friendly open-source codebase for running LUV and training segmentation networks.
3) Publicly-available, annotated datasets for fabric corner keypoints (3640 labeled images), cable segmentation masks (486 labeled images), and needle segmentation masks (1364 labeled images).
4) Experimental results evaluating the LUV-trained segmentation results in terms of flexibility and performance on 3 real-world robot perception tasks including locating fabric keypoints, cables and surgical needles segmentation. We report intersection over union (IOU) metrics showing LUV produces accurate labels, enables 83% task success on folding towels from corner detections, and localizes needles within 1.7mm of human labels.

II. Related Work

A. Semantic Segmentation

Semantic segmentation is a well-studied field in computer vision with significant advances in the past decade due to the emergence of large labeled datasets like COCO, PASCAL VOC, CityScapes [4, 9, 23], and the development of segmentation architectures like fully-convolutional networks (FCNs) [24], U-Nets [33], and region-proposal networks [32]. Training these networks relies on a large dataset of images with pixel-wise annotations of objects. Previous work accomplishes this with cloud-based human image labeling, which distributes the task of labeling data to human laborers on platforms like Amazon Mechanical Turk or Scale [1, 34]. This method, though effective, suffers from inconsistent label quality, difficulty in specifying labels in ambiguous situations, and cumbersome oversight processes to filter low quality labels. In addition, for involved tasks like image segmentation, the recommended price on Turk is $0.82 [1] per image.

B. Self-supervised Robot Data Collection and Labeling

To alleviate the need for explicit human labels, many prior works automate data collection and labeling by leveraging specific structure available in the task. This is commonly applied when training dynamics models that predict the resulting state after an action, both on images [7, 10, 16, 46] and lower-dimensional state such as keypoints [27]. Recently, self-supervision has been applied in imitation learning to obtain ground-truth action labels for image-based policies by manually resetting the robot to a goal or known configuration, perturbing the end effector by a known displacement, and using the displacement with the initial pose to compute an action label [8, 25, 45]. Self-supervision is also a popular technique in reinforcement learning when automatic resets are available. Kalashnikov *et al.* [20] and Pinto *et al.* [30] use the result of autonomously explored robot grasps to supervise a grasp quality estimator, making it possible for them to collect 580K and 50K physical grasps respectively. LUV makes state/label estimation possible in situations where autonomous labels were previously difficult to generate and can be applied to extend the above self-supervision techniques. LUV is similar to Qian *et al.* [31], who use visible markers to label images for a network that predicts cloth features from depth images alone. In constrast to this work, LUV can be used on pure RGB images, which is useful in tasks such as needle segmentation where active depth sensors tend to fail [45].

C. Fluorescent Marking Technology

Fluorescent markings are also commonly used in non-robotics applications to track and identify target objects. In medicine, near-infrared (NIR) fluorescence is used for cancer treatment [35]. In water treatment, fluorescence spectroscopy is applied to identify fouling agents and monitor wastewater quality [2]. In robotics, UV-fluorescent paints have been used for tracking cloth state [39]. In contrast, we do not use UV-fluorescence at execution time, and instead use the fluorescent markings to label training data. The most related work to this is Takahashi *et al.* [43], which marks fluid, powder, and clothing with UV-fluorescent dye and paint, then uses a custom-designed strobing LED system to collect standard images and their corresponding annotations. This paper extends their approach to other materials, makes it more accessible with off-the-shelf parts to collect data in a self-supervised way, and shows results on practical robotics tasks.

III. Labels from UltraViolet (LUV)

In this section, we present Labels from UltraViolet (LUV), a framework for generating image annotations in manipulation domains without human labeling.

A. Framework Overview

LUV is comprised of two phases: **training** and **execution**.

1) Training: During training, relevant keypoints and segmentation masks are painted with transparent UV fluorescent paint, that are nearly invisible in natural lighting but brightly light up in different colors under UV radiation. The robot collects a dataset of paired images in the workspace: $\mathcal{D}_{\text{train}} = \left\{(I_{i,\text{std}}, I_{i,\text{uv}})_{i=1}^{N_{\text{train}}}\right\}$. Each standard RGB image $I_{i,\text{std}} \in \mathbb{R}^{H \times W \times 3}$ is taken under standard workspace lighting conditions. For the UV RGB images, $I_{i,\text{uv}} \in \mathbb{R}^{H \times W \times 3}$, the workspace is illuminated by ultraviolet spectrum LED lamps. The workspace is otherwise unmodified between the paired images. The fluorescence is used to extract segmentation

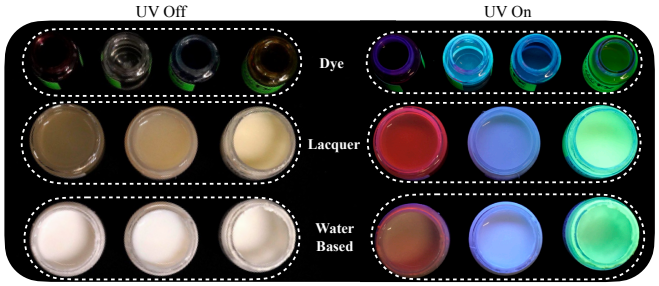

Fig. 2: **UV Paint Types**: We consider three types of UV-fluorescent paint in this paper: dyes (top), lacquer-based paint (middle), and water-based paints (bottom). We describe their properties and painting techniques in Section III-B. Most paints dry almost clear.

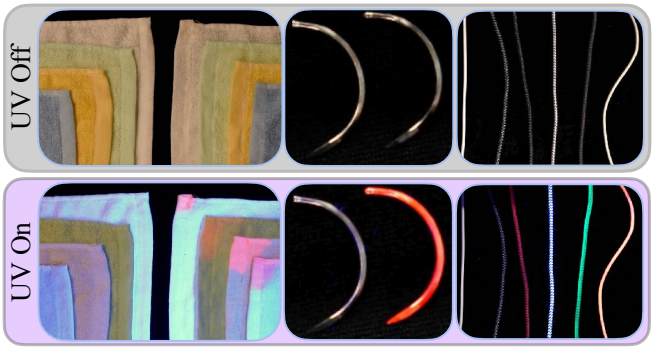

Fig. 3: **UV Paint Transparency**: The top row of this figure contains both painted and unpainted towels, surgical needles, and household charging cables under standard lighting. The left objects in each image are unpainted and the right objects are painted with transparent UV paint. Under standard lighting, the painted objects are difficult to visually distinguish from the unpainted objects, without careful inspection. Under UV radiation (bottom row), the painted objects distinctly fluoresce based on the color of the paint used.

masks and keypoints from the UV images, which are used as training labels for the standard images. A learning-based perception model f_θ is trained on the labeled dataset.

2) Execution: During execution, the perception model f_θ is evaluated only on images under standard workspace lighting conditions containing unpainted objects.

B. UV Fluorescent Paint

We experiment with three types of fluorescent marking substances (Figure 2) and describe the most successful techniques for applying them, as well as their robustness and surface finish properties.

1) Lacquer-based paint: This is a viscous paint consisting of fluorescent powder dissolved in a lacquer which dries clear and glossy. The red and blue is completely dissolved, yielding a glassy finish, while the green is only partially dissolved with some suspended particles. The green paint thus leaves behind a faint white powder when dried [14]. This paint can be thinned with standard lacquer thinner, making it more suitable for applying to deformables without stiffening.

2) Dye: This substance is a watery staining fluid which works well on absorbent materials. There is one type which is completely transparent under visible light and fluoresces blue, and a variety of other fluorescent colors which have color under visible light. The clear variant is completely invisible under standard lighting, while the colored dyes are only invisible on materials colored similarly. On light colored materials, these dyes can be diluted to further minimize staining and save cost [6].

3) Water-based paint: This paint is acrylic and dries translucent. It is invisible on lighter colored materials but leaves a faint milky residue on darker materials [14].

C. Test Material Properties

This section describes some important factors to consider when choosing a marking type for a new material.

1) Natural fluorescence: Some materials, particularly white colored papers and cloths, exhibit natural blue fluorescence, prohibiting the use of blue markings on them. Other color fluorescent markings will work on these materials, such as the white towel in our experiment, though the blue fluorescence will shift the color of marks when exposed to UV light. Avoiding materials which naturally fluoresce in the scene is thus desirable for ease of marking and post-processing.

2) Fibrous materials: Diluted lacquer paint and dyes are particularly well suited for cloth, whereas undiluted lacquer or water-based paint should be used for solid objects.

3) Color: Dark objects result in weaker fluorescence due to light absorption, though in our experience this is typically only significant with near-black materials and cloth.

4) Luster: Matte materials are better suited to water-based paint. Shiny objects match the surface finish of lacquer paint.

D. UV Lighting

We describe the UV lights used to illuminate the setup and the smartplugs which automatically toggle them.

1) UV Lights: LUV uses 365nm LED UV floodlights to bathe the workspace in light from multiple angles to minimize shadows which fragment segmentation masks. Tube-fluorescent bulbs are *not* usable for this task because they cannot be switched rapidly, while LEDs nearly instantly trigger. Shorter wavelength UV lights yield brighter fluorescence, hence the choice of 365nm light is important over other available 405nm LEDs. LUV can work in settings with strong ambient room lighting, as in the cable segmentation and towel smoothing datasets, or in controlled lighting settings where room lights can be switched off, as in needle segmentation. Toggleable ambient lights optionally make fluorescent labels stand out more from the background.

2) Toggling: We use an off-the-shelf smart plug which plugs into any standard wall socket, connects to a local network, and is controllable from any device on the same network via a Python interface. This allows programmatically switching lights on and off during data collection, and is a scalable solution for any number of lights by plugging a power adapter into the smart plug.

E. Mask Generation

To generate masks, the UV lights are turned on, and if available the ambient white lights turned off. The camera exposure for each sample is found by manually sweeping

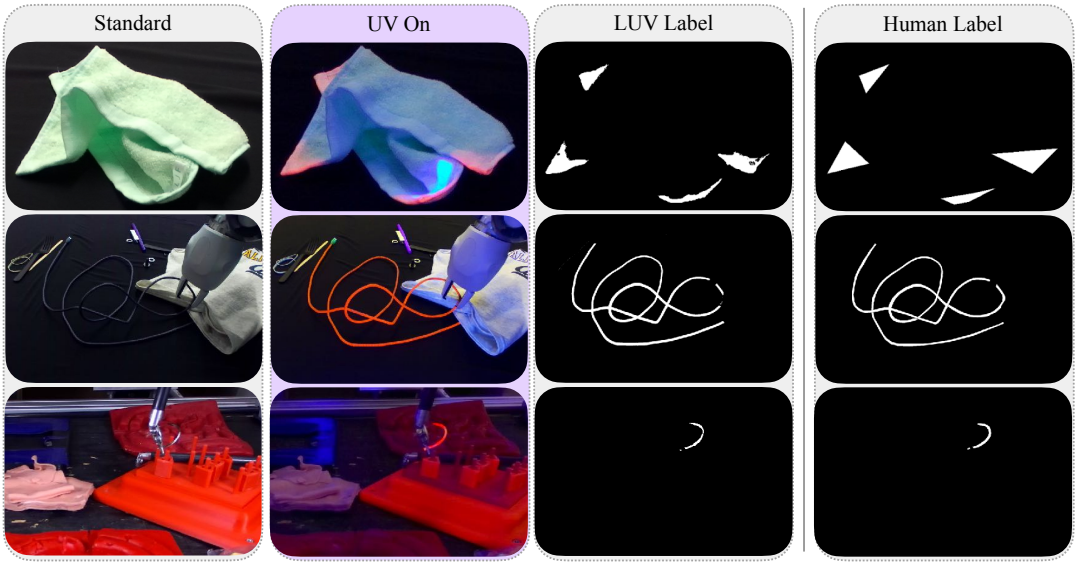

Fig. 4: **LUV Data Collection**: We consider three tasks from prior robot manipulation literature: cloth corner detection, cable segmentation, and needle segmentation. To collect a labeled datapoint, LUV collects an image in standard lighting and then collects an image of the same scene under UV lighting. Then, color segmentation is used to extract the relevant annotations for the image. HSV filtering described in Section III-E is able to extract the red fluorescence of the needle without capturing any of the other red objects in the scene. We quantitatively compare the consistency of the masks to human labels (right column) in Table **??**.

exposures and selecting the exposure yielding clearest label colors. We use the Zed M stereo camera, which provides a software interface to programmatically control these settings. White-balance is held constant during data collection, and HSV color thresholds are hand-picked from a sample image. For our tasks this calibration process takes only a few minutes, though it could be streamlined by implementing a user interface for automatically picking exposure and thresholds. Figure 4 shows examples of extracted masks.

For scenes with both dark and light painted materials, multiple exposures can be captured and post-processed with HDR [28] to retrieve well exposed labels for all colors.

F. Parts List and Cost

The total 1-time cost of setting up LUV for indoor ambient lighting at the time of paper submission is $282. Based on Amazon's recommended price of $0.82 per semantic segmentation label on Amazon Mechanical Turk, and using 2 labels per image based on their recommendation for quality [1], this breaks even with Turk at 167 labeled images. In contrast, several of the datasets generated in this work contain well over 1000 labeled images (Section IV), making the cost of LUV more than 5x less expensive.

Parts needed to set up a minimal working system are

1) Lights: Everbeam 100W LED 365nm floodlight, available on Amazon for $88 each.
2) Smart Plug: Kasa Smart Plug HS103P2, available as a 2-pack on Amazon for $18.
3) Camera: Any existing RGB camera will work, however for best results it should have exposure control and manual white-balance options.
4) Fluorescent Marking: To get started, we recommend beginning with lacquer based paints, being the most versatile and invisible. Our paint is sourced from the company "GloEffex" under the product name "Transparent UV Paint" [14].

IV. Experiments

We evaluate LUV on a set of perception tasks commonly studied in robot manipulation, but LUV can in principle be applied to other tasks such as keypoint detection. Experiments are designed to evaluate the label quality, data collection rate, and flexibility of LUV compared to human labeling. Due to the much faster rate of data collection, we are able to increase the difficulty of several tasks compared to prior work by considering more visually challenging scenes including distractor objects. In all experiments, RGB data is collected using a Zed M stereo camera. We use a U-Net [33] in the towel corner detection task, and we use a ResNet50-FCN [24] for the cable and needle tasks.

Evaluation Metrics: In the needle and cable segmentation tasks, we compare the labeling quality and throughput of LUV to human labeling. We report the average intersection over union (**IOU**) between masks from LUV and from a human labeler on a set of 10 training images. We also report the average seconds per label (**SPL**) for a human labeler and for LUV to annotate each of these images. We evaluate the quality of the learned network on an unseen test set of images in each domain containing only unpainted objects by reporting the intersection over union (**IOU**) of the predictions compared to human labels on these test images.

A. Towel Corner Detection for Smoothing and Folding

Predicting task-relevant keypoints using fully-convolutional neural networks is a popular technique in deformable manipulation applications such as fabric smoothing [12, 36], t-shirt folding [12, 21], and cable untangling [15, 40, 44]. Seita *et al.* [36] and Ganapathi

et al. [12] both predict keypoints corresponding to the corners of a rectangular towel to implement a smoothing policy that iteratively pulls identified corners away from the towel. Both methods are trained on large datasets of simulation data. In this experiment, we collect a dataset of real images containing diverse colors of towels in different configurations with a label on each corner of the towels. We train a neural network to predict towel corners and later use it in an algorithm that smooths and folds towels.

1) Experimental Setup and Assumptions: The workspace contains an ABB YuMi robot, facing a tabletop with a black tablecloth. We sample towels to place in the workspace from a set of four, monochromatic, rectangular towels. We assume, for this task, that no other objects are in the scene.

2) Task Definition: The goal of this task is to predict masks corresponding to the corners of a towel in the workspace from input images. The predicted masks are used by a heuristic algorithm to smooth and fold the towel. We attempt square double-folds and consider towels "folded" if the final state of the towel can be pinched at the innermost folded corner and shaken without disrupting the folds in the towel, meaning 4 layers of cloth closely occupy the corner.

3) Data Collection: Self-supervised data collection is a popular technique for training fabric manipulation policies [10, 16]. Data for this task is collected autonomously by color thresholding the towel from the black background, picking it up from a random point along the border, shaking it in the air, and dropping it. Because this process often biases towards crumpled states, we manually place the fabric in smoother configurations and collect a small set of images in more orderly configurations as well. Data is collected with the ABB YuMi robot. The training dataset has 3640 images.

4) Smoothing and Folding Algorithm: Using the corner outputs from the network, we implemented a heuristic algorithm, to first smooth a crumpled towel then fold the smoothed towel, shown in Fig. 7. The algorithm repeats *smoothing actions* until it detects the towel is smoothed. During each action, if no corner is detected, the robot randomly resets the cloth by grasping at a random point, shaking and dropping it. If only one corner is detected, the robot grasps the visible corner, shakes it slightly and drags it across the table to spread other corners out. If more than one corner are detected, the robot grasps the pair of corners closest to each other, lifts them up and flings them forwards and back, flattening the towel. This process terminates when 4 corners are detected whose pairwise distances match the size of the towel within a standard deviation of their mean.

After the towel is smoothed, as shown in Fig. 7 panel 3, the locations of all 4 corners are measured and used to execute an open-loop folding motion. First, the robot grasps the two corners near the robot and puts them on top of the further two corners. Then, a single arm grasps the short folded segment and places it to match the antipodal one, completing the fold.

We use the depth output from the Zed stereo camera to retrieve the 3D positions of the corners by taking the median of deprojected depth points inside the network output mask, and execute grasps using a fixed gripper orientation.

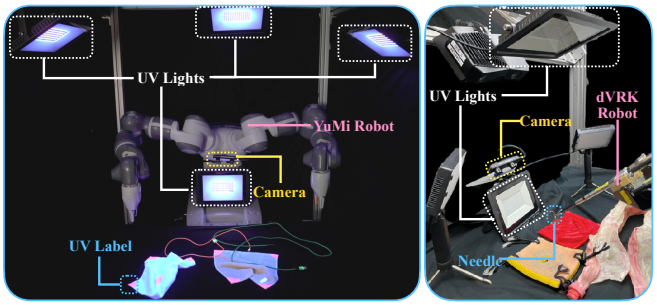

Fig. 5: **Experimental Setups:** We use LUV on two robot setups. **Left:** The first, a bimanual YuMi robot, consists of 4 UV lights oriented at different angles to mazimize UV coverage, with a camera mounted between the arms to minimize arm occlusions. All 4 UV lights are toggled with the same smart plug. **Right:** The second, a dVRK surgical robot, has 2 UV lights and 2 visible LED lamps which are each controlled by separate smartplugs. The UV lights are turned on when the visible lamps are turned off and vice versa.

5) Results: We evaluate the algorithm on towels in the train set, with random initializations as in autonomous data collection. We execute at most 10 smoothing actions before considering the rollout a failure. Smoothing is successful if it terminates autonomously and proceeds to folding.

Results are reported in table **??**. Smoothing succeeds on average 92% of rollouts across all towels, and folding on average 83%. The majority of failure cases are from manipulation challenges and the heuristic algorithm, such as timeouts from repeated grasp failures, or looping because the algorithm grasps diagonal corners over and over, rather than corners detection failure.

B. Cable Segmentation

Prior work in cable manipulation [15, 40, 44] often assumes that the cable is visually distinguishable from the background via color segmentation. But cables in household or industrial settings may not be chromatically distinct from background objects. This motivates a learning-based approach to predict cable segmentation masks from images. However, labeling cable segmentation masks is extremely tedious due to the complexity of cable configurations. We apply LUV to the task of cable segmentation, and include configurations with distractor objects and multiple cables in the scene. Generating these complex segmentation labels via LUV takes less than 178 ms per image (Table **??**).

1) Experimental Setup and Assumptions: The workspace contains a bilateral ABB Yumi robot. We sample cables to place in the scene from a set of 2 micro-USB to USB cables and a lightning to USB cable. We found that one of the micro-USB cables is naturally fluoresces blue under UV radiation without any painting, so we use this cable in both the training and execution images since we do not paint it. For the other cables, we use a painted version for training images and an unpainted version for execution images. To increase both the difficulty of the task and robustness of the model, we place unpainted distractor objects in the workspace. Several of the items are selected to contain reflections and colors similar to the cables used. At execution

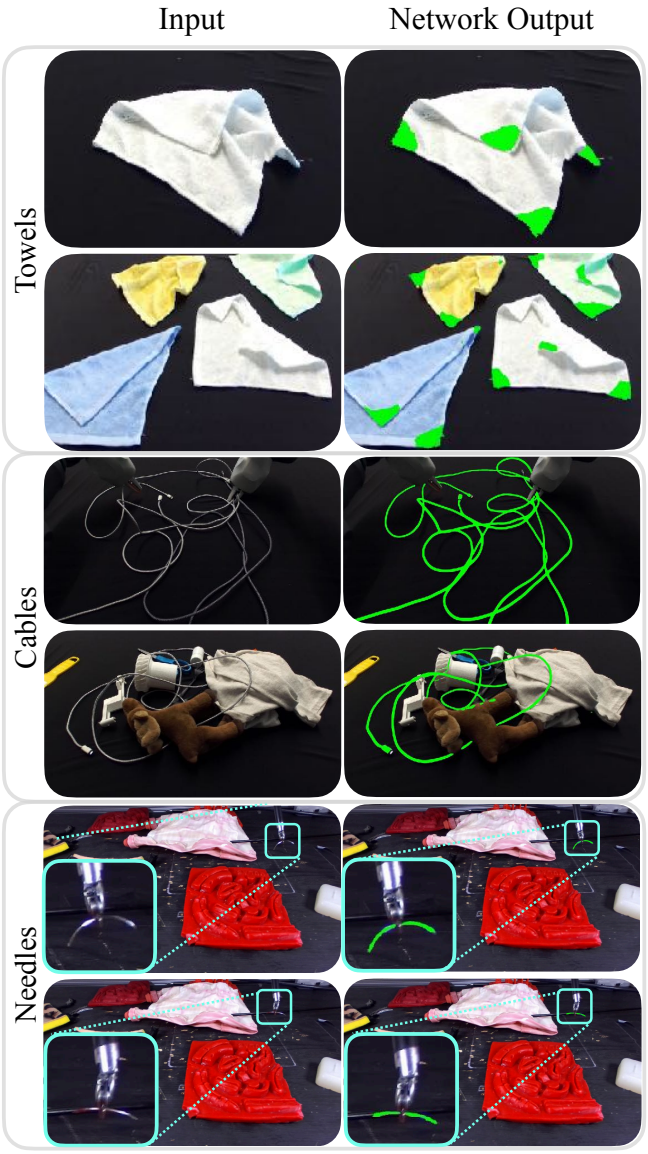

Fig. 6: **Network Predictions on Unseen, Test Images**: We evaluate the trained networks for each of the tasks on images unseen during training. The input images are in the left column, and the right column depicts the input images with the predicted segmentation masks overlaid in green. To test generalization, some of the towel images contain multiple towels, even though all of the training images only had a single towel. The leftmost cable test image contains two cables, even though all of the training images only contained a single cable.

time, we introduce several novel distractor objects unseen in training images.

2) Task Definition: The goal of this task is to predict a semantic segmentation mask $I_{\text{seg,cable}}$ corresponding to all of the cables in an input image of the workspace.

3) Data Collection: We collect data separately for each of the painted cables. We place each cable in the scene and manually randomize its configuration, knot structure, and position of the YuMi arms. We also randomly place distractor items in the scene. We collect 486 labeled images.

4) Results: Due to the visual complexity of these scenes, collecting human annotated images was extremely time consuming, with each image taking an average of 446 seconds to

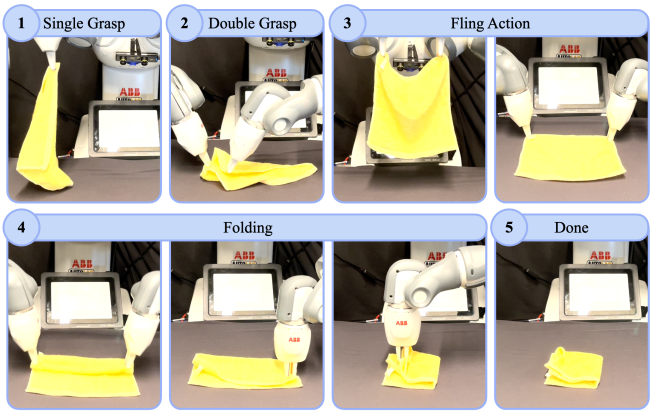

Fig. 7: **Towel Folding:** During smoothing, we use corner predictions from a trained network to implement a heuristic smoothing algorithm. If one corner is visible, the robot drags the towel sideways (**1**) from this corner to increase others' visibility. If multiple are visible, the robot grabs the two nearest corners (**2**) and flings (**3**) across the table. When 4 detected corners are arranged in a square, the robot executes a folding motion using corner positions to compute grasp positions (**4**), leading to a neatly folded final product (**5**). Detailed results are shown in table **??**.

label (Table **??**). LUV takes an average of 0.178 seconds to annotate each image, which is 2511 *times faster*. Labeling all of the image in the dataset with LUV takes about 87 seconds *in a single thread*, whereas we estimate from Table **??** that labeling all of the collected images would take over $446 \times 486 = 60$ hours for a human annotator. We find that the mean IOU between LUV labels and human labels on 10 training images is 0.787 (Table **??**) and the mean IOU between the LUV-trained segmentation network and human labels on a test set of 10 images is 0.755 (Table **??**).

C. Needle Segmentation

Segmenting surgical needles in images is a common perception task in surgical robotics research. Some prior works rely on painting the needle and performing color segmentation [3, 38, 45] or assuming that the rest of the scene is visually distinct from the needle [42]. Recent work uses fully-convolutional neural networks to predict these masks, using a combination of simulation and real data [45], but restrict their task to black backgrounds. In this task, we apply LUV to needle segmentation in the presence of tissue phantoms and distractor objects in the scene.

1) Experimental Setup and Assumptions: We collect data in a workspace with a bilateral da Vinci Research Kit surgical robot and an inclined Zed M stereo camera. We collect training data using a set of 2 surgical needles and test data using unpainted versions of these needles. In contrast to prior work, which considers a pristine black background for color segmentation [45], we increase the difficulty and realism of the task by randomly placing surgical phantoms, training equipment, and tools in the scene.

2) Task Definition: The goal of this task is to predict a semantic segmentation mask for all surgical needles.

3) Data Collection: Unlike [45], we exclusively train methods on real data, and we use their self-supervised data collection policy to move the needle around the workspace

	IOU	SPL (human)	SPL (LUV)
Towel Corner Detection (Train)	N/A	22.5	**0.125**
Cable Segmentation (Train)	0.787	446	**0.178**
Needle Segmentation (Train)	0.683	40	**0.221**
Cable Segmentation (Execution)	0.755	N/A	N/A
Needle Segmentation (Execution)	0.666	N/A	N/A

TABLE I: **Labeling Technique Comparison:** We evaluate LUV on a set of 3 robot perception tasks. **Top half:** We compare the consistency of the UV training labels with human labels as a measure of label quality by comparing their intersection over union (IOU). We report the seconds per label (S.P.L) for both a human labeler and LUV. On the segmentation tasks, we observe that the training masks for LUV have an IOU of 0.787 and 0.683 with respect to human labeled masks. Because the cables and needles are very thin, small discrepancies can significantly impact the IOU score negatively, and the labels were also challenging for a human to label. On the needle task, we quantify label quality by using them for pose estimation in Table **??**. We observe that LUV takes 180-2511$\times$ less time than the human to label images. Labeling the entire cable dataset by hand would take approximately 60 hours, whereas it takes LUV 87 seconds in a single-thread. **Bottom half:** We evaluate the models trained with LUV labels on unpainted test images and compare the predictions to human labels. We find that the predictions have an average intersection over union (IOU) of 0.755 and 0.666 on the two tasks. Because cables and needles have thin segmentation masks, small discrepancies with respect to human labels can lead to large negative drops in IOU.

Towel	No. Smooth Action	Smooth Success	Fold Success
Blue	4.4 ± 1.95	83%	67%
Green	5.0 ± 3.16	100%	100%
White	1.8 ± 0.41	100%	100%
Yellow	4.0 ± 2.91	83%	67%
Average	3.7 ± 2.52	92%	83%

TABLE II: **Smoothing and Folding Results**. For each towel, 6 trials are conducted with random initial states leading to 24 trials in total. Mean and standard deviation of number of smoothing actions, smoothing and folding success rate are reported.

when grasped by the robot's end effector. We periodically insert, remove, and move surgical tissue phantoms in the background like silicone suture practice pads and simulated gut. We use the da Vinci Research Kit for data collection. The needle dataset consists of 1364 images, with infrequent human interventions to periodically change the poses of the needle in the gripper and the background objects.

4) Results: We compare the annotations generated by LUV to human annotations on a set of 38 training images. The mean IOU between the segmentation masks in the two sets is 0.683. We train a segmentation network whose prediction has a mean IOU of 0.666 with respect to human annotations on an unpainted test set of 20 images. While this seems relatively low, this is due to the very thin profile of needles, and slight variations in the human and LUV annotations can significantly affect IOU. The masks are qualitatively very similar as shown in Fig. 4, and we quantify this by running the needle pose reconstruction algorithm from Wilcox *et al.* [45] on stereo images in both the training and test set. We limit test images to those where both tips of the needle are visible. The resulting average pose error between labels from LUV and humans is 1.7mm and 6.9°, and the pose error on the test set between network outputs and human labels is 1.7mm and 8.8°.

	Position Difference(mm)	Rotation Difference
Training Labels	1.7mm	6.9°
Test Predictions	1.7mm	8.8°

TABLE III: **LUV Needle Pose Estimate Consistency. Top Row:** We use the training needle masks (Training Labels) generated by the UV labels to estimate the needle's pose using the 3D needle pose reconstruction algorithm from Wilcox *et al.* [45]. We compare the pose to the pose generated by human labels of the training images. We find that the pose estimate is within 1.7mm and 6.9° the poses generated by human labels on average. In this test, we ensure that the images do not contain the needle in pathological cases, as described in Wilcox *et al.* [45]. **Bottom Row:** We evaluate the trained network predictions on test images in the same way, and find that the resulting average pose estimate is within 1.7mm and 8.8° of the pose generated by the human labeled segmasks.

V. Discussion

We present Labels from UltraViolet (LUV), a framework for collecting segmentation labels and keypoints without human labeling. LUV can be applied to augment existing self-supervised dataset collection techniques in robot manipulation domains. In future work, we hope to investigate other applications of LUV, such as using fluorescent markers to obtain object poses, detecting garment edges, or using UV markers for online dense reward assignment in RL tasks. In addition, LUV could be extended to produce a larger diverse dataset of fabric annotations on RGB images.

Acknowledgments

This research was performed at the AUTOLAB at UC Berkeley in affiliation with the Berkeley AI Research (BAIR) Lab, the CITRIS "People and Robots" (CPAR) Initiative, the Real-Time Intelligent Secure Execution (RISE) Lab and UC Berkeley's Center for Automation and Learning for Medical Robotics (Cal-MR). This work is supported in part by donations from Intuitive Surgical, Google, Siemens, Autodesk, Bosch, Toyota Research Institute, Honda, Intel, Hewlett-Packard and by equipment grants from PhotoNeo and NVidia. We thank Adam Lau for his photography and Kishore Srinivas for assistance with data labelling.

References

[1] *Amazon sagemaker data labeling pricing*, `https://aws.amazon.com/sagemaker/data-labeling/pricing/?nc=sn&loc=3`.

[2] E. M. Carstea, J. Bridgeman, A. Baker, and D. M. Reynolds, "Fluorescence spectroscopy for wastewater monitoring: A review," *Water research*, vol. 95, pp. 205–219, 2016.

[3] Z.-Y. Chiu, F. Richter, E. K. Funk, R. K. Orosco, and M. C. Yip, "Bimanual regrasping for suture needles using reinforcement learning for rapid motion planning," *arXiv:2011.04813*, 2020.

[4] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[5] M. Danielczuk, M. Matl, S. Gupta, A. Li, A. Lee, J. Mahler, and K. Goldberg, "Segmenting unknown 3d objects from real depth images using mask r-cnn trained on synthetic data," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019.

[6] *Dark light fx*, `https://darklightfx.com/`.

[7] S. Dasari, F. Ebert, S. Tian, S. Nair, B. Bucher, K. Schmeckpeper, S. Singh, S. Levine, and C. Finn, "Robonet: Large-scale multi-robot learning," *arXiv preprint arXiv:1910.11215*, 2019.

[8] N. Di Palo and E. Johns, "Learning multi-stage tasks with one demonstration via self-replay," in *Conference on Robot Learning*, PMLR, 2022, pp. 1180–1189.

[9] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, 2010.

[10] C. Finn and S. Levine, "Deep visual foresight for planning robot motion," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2017, pp. 2786–2793.

[11] P. R. Florence, L. Manuelli, and R. Tedrake, "Dense object nets: Learning dense visual object descriptors by and for robotic manipulation," *arXiv preprint arXiv:1806.08756*, 2018.

[12] A. Ganapathi, P. Sundaresan, B. Thananjeyan, A. Balakrishna, D. Seita, J. Grannen, M. Hwang, R. Hoque, J. E. Gonzalez, N. Jamali, *et al.*, "Learning dense visual correspondences in simulation to smooth and fold real fabrics," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2021, pp. 11 515–11 522.

[13] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[14] *Glo-effex: Glow in the dark and uv reactive paints*, `https://www.gloeffex.com/`.

[15] J. Grannen, P. Sundaresan, B. Thananjeyan, J. Ichnowski, A. Balakrishna, M. Hwang, V. Viswanath, M. Laskey, J. E. Gonzalez, and K. Goldberg, "Untangling dense knots by learning task-relevant keypoints," *arXiv preprint arXiv:2011.04999*, 2020.

[16] R. Hoque, D. Seita, A. Balakrishna, A. Ganapathi, A. K. Tanwani, N. Jamali, K. Yamane, S. Iba, and K. Goldberg, "Visuospatial foresight for multi-step, multi-task fabric manipulation," *arXiv preprint arXiv:2003.09044*, 2020.

[17] H. Huang, M. Danielczuk, C. M. Kim, L. Fu, Z. Tam, J. Ichnowski, A. Angelova, B. Ichter, and K. Goldberg, *Mechanical search on shelves using a novel "bluction" tool*, 2022. arXiv: `2201.08968`.

[18] H. Huang, M. Dominguez-Kuhne, V. Satish, M. Danielczuk, K. Sanders, J. Ichnowski, A. Lee, A. Angelova, V. O. Vanhoucke, and K. Goldberg, "Mechanical search on shelves using lax-ray: Lateral access x-ray," 2021.

[19] J. J. Ji, S. Krishnan, V. Patel, D. Fer, and K. Goldberg, "Learning 2d surgical camera motion from demonstrations," in *2018 IEEE 14th International Conference on Automation Science and Engineering (CASE)*, IEEE, 2018, pp. 35–42.

[20] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, *et al.*, "Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation," *arXiv preprint arXiv:1806.10293*, 2018.

[21] T. Kollar, M. Laskey, K. Stone, B. Thananjeyan, and M. Tjersland, "Simnet: Enabling robust unknown object manipulation from pure synthetic data via stereo," *arXiv preprint arXiv:2106.16118*, 2021.

[22] V. Lim, H. Huang, L. Y. Chen, J. Wang, J. Ichnowski, D. Seita, M. Laskey, and K. Goldberg, *Planar robot casting with real2sim2real self-supervised learning*, 2021. arXiv: `2111.04814 [cs.RO]`.

[23] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, 2014.

[24] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[25] D. Ma, S. Dong, and A. Rodriguez, "Extrinsic contact sensing with relative-motion tracking from distributed tactile measurements," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2021, pp. 11 262–11 268.

[26] J. Mahler, M. Matl, V. Satish, M. Danielczuk, B. DeRose, S. McKinley, and K. Goldberg, "Learning ambidextrous robot grasping policies," *Science Robotics*, vol. 4, no. 26, eaau4984, 2019.

[27] L. Manuelli, Y. Li, P. Florence, and R. Tedrake, "Keypoints into the future: Self-supervised correspondence in model-based reinforcement learning," *arXiv preprint arXiv:2009.05085*, 2020.

[28] T. Mertens, J. Kautz, and F. Van Reeth, "Exposure fusion," in *15th Pacific Conference on Computer Graphics and Applications (PG'07)*, IEEE, 2007, pp. 382–390.

[29] S. Paradis, M. Hwang, B. Thananjeyan, J. Ichnowski, D. Seita, D. Fer, T. Low, J. E. Gonzalez, and K. Goldberg, in *Intermittent Visual Servoing: Efficiently Learning Policies Robust to Instrument Changes for High-precision Surgical Manipulation*, IEEE International Conference on Robotics and Automation (ICRA), 2021.

[30] L. Pinto and A. Gupta, "Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours," in *2016 IEEE international conference on robotics and automation (ICRA)*, 2016.

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

[32] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[33] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, Springer, 2015, pp. 234–241.

[34] *Scale ai - customers*, `https://scale.com/customers`.

[35] B. E. Schaafsma, J. S. D. Mieog, M. Hutteman, J. R. Van der Vorst, P. J. Kuppen, C. W. Löwik, J. V. Frangioni, C. J. Van de Velde, and A. L. Vahrmeijer, "The clinical use of indocyanine green as a near-infrared fluorescent contrast agent for image-guided oncologic surgery," *Journal of surgical oncology*, vol. 104, no. 3, 2011.

[36] D. Seita, A. Ganapathi, R. Hoque, M. Hwang, E. Cen, A. K. Tanwani, A. Balakrishna, B. Thananjeyan, J. Ichnowski, N. Jamali, K. Yamane, S. Iba, J. Canny, and K. Goldberg, in *Deep Imitation Learning of Sequential Fabric Smoothing From an Algorithmic Supervisor*, IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2020.

[37] D. Seita, N. Jamali, M. Laskey, A. K. Tanwani, R. Berenstein, P. Baskaran, S. Iba, J. Canny, and K. Goldberg, "Deep transfer learning of pick points on fabric for robot bed-making," *arXiv preprint arXiv:1809.09810*, 2018.

[38] S. Sen, A. Garg, D. V. Gealy, S. McKinley, Y. Jen, and K. Goldberg, in *Automating Multiple-Throw Multilateral Surgical Suturing with a Mechanical Needle Guide and Sequential Convex Optimization*, IEEE International Conference on Robotics and Automation (ICRA), 2016.

[39] E. Siyu, "Evaluation of visible and invisible fiducial markers for clothing tracking," *Electrical Engineering and Computer Sciences, University of California at Berkeley*, 2012.

[40] P. Sundaresan, J. Grannen, B. Thananjeyan, A. Balakrishna, J. Ichnowski, E. Novoseller, M. Hwang, M. Laskey, J. E. Gonzalez, and K. Goldberg, "Untangling dense non-planar knots by learning manipulation features and recovery policies," *arXiv preprint arXiv:2107.08942*, 2021.

[41] P. Sundaresan, J. Grannen, B. Thananjeyan, A. Balakrishna, M. Laskey, K. Stone, J. E. Gonzalez, and K. Goldberg, "Learning rope manipulation policies using dense object descriptors trained on synthetic depth data," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2020, pp. 9411–9418.

[42] P. Sundaresan, B. Thananjeyan, J. Chiu, D. Fer, and K. Goldberg, "Automated extraction of surgical needles from tissue phantoms," in *2019 IEEE 15th International Conference on Automation Science and Engineering (CASE)*, IEEE, 2019, pp. 170–177.

[43] K. Takahashi and K. Yonekura, "Invisible marker: Automatic annotation of segmentation masks for object manipulation," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 8431–8438.

[44] V. Viswanath, J. Grannen, P. Sundaresan, B. Thananjeyan, A. Balakrishna, E. Novoseller, J. Ichnowski, M. Laskey, J. E. Gonzalez, and K. Goldberg, "Disentangling dense multi-cable knots," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2021, pp. 3731–3738.

[45] A. Wilcox, J. Kerr, B. Thananjeyan, J. Ichnowski, M. Hwang, S. Paradis, D. Fer, and K. Goldberg, "Learning to localize, grasp, and hand over unmodified surgical needles," *IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[46] A. Xie, F. Ebert, S. Levine, and C. Finn, "Improvisation through physical understanding: Using novel objects as tools with visual foresight," *arXiv preprint arXiv:1904.05538*, 2019.

[47] M. Yan, Y. Zhu, N. Jin, and J. Bohg, "Self-supervised learning of state estimation for manipulating deformable linear objects," *IEEE robotics and automation letters*, vol. 5, no. 2, pp. 2372–2379, 2020.

[48] H. Zhang, J. Ichnowski, D. Seita, J. Wang, H. Huang, and K. Goldberg, "Robots of the lost arc: Self-supervised learning to dynamically manipulate fixed-endpoint cables," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2021, pp. 4560–4567.