
Touch and Go: Learning from Human-Collected Vision and Touch

Fengyu Yang^{1*} Chenyang Ma^{1*} Jiacheng Zhang¹
Jing Zhu¹ Wenzhen Yuan² Andrew Owens¹

¹University of Michigan ²Carnegie Mellon University

<https://touch-and-go.github.io>

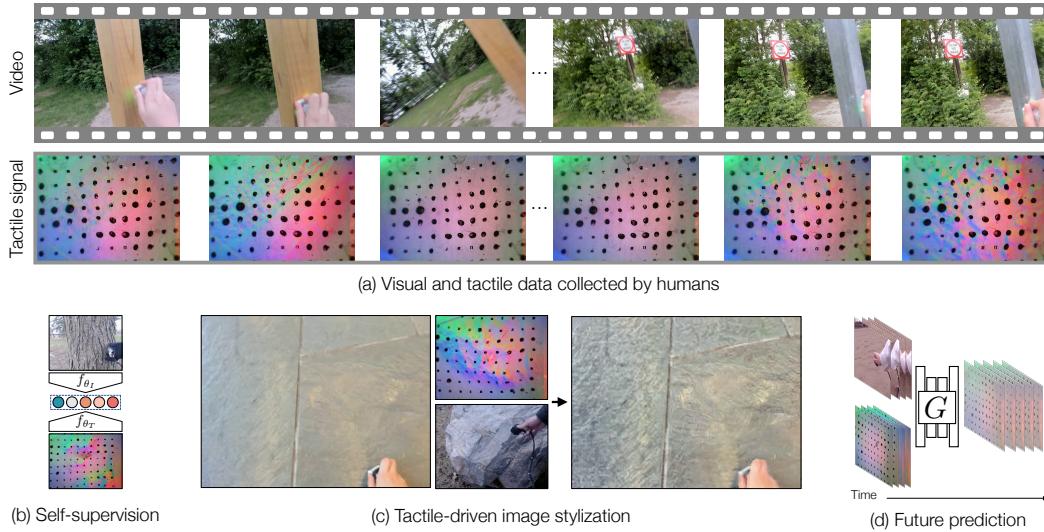


Figure 1: **The *Touch and Go* dataset.** We collect a dataset of real-world visual and touch data. (a) Humans walk through a large number of scenes, probing objects around them with a touch sensor and recording video. We apply this dataset to: (b) learning tactile features through self-supervision, by associating touch with sight, (c) manipulating an image to match the tactile signal (e.g., restyling a smooth surface to match the tactile signal for a rough rock, whose photo we show for reference), (d) predicting future frames of a tactile signal from visuo-tactile inputs.

Abstract

1 The ability to associate sight with touch is essential for understanding material
2 properties, and for physically interacting with the world. Learning these corre-
3 lations, however, has proven challenging, since existing datasets have not captured
4 the full diversity of these modalities. To address this shortcoming, we propose a
5 dataset for multimodal visuo-tactile learning called *Touch and Go*, in which human
6 data collectors probe objects in natural environments with tactile sensors, while
7 recording egocentric video. The objects and scenes in our dataset are significantly
8 more diverse than prior efforts, making the data well-suited to tasks that involve
9 understanding material properties and physical interactions in the wild. To demon-
10 strate our dataset’s effectiveness, we successfully apply it to a variety of tasks: 1)
11 self-supervised visuo-tactile feature learning, 2) tactile-driven image stylization,
12 i.e., making the visual appearance of an object more consistent with a given tactile
13 signal, and 3) predicting future frames of a tactile signal from visuo-tactile inputs.

*The first two authors contributed equally to this paper.

14 **1 Introduction**

15 As humans, our ability to correlate touch with sight is an essential component of understanding
16 the physical properties of the objects around us, and of interacting with the world. While recent
17 advances in other areas of multimodal learning have been fueled by large datasets, the difficulty of
18 collecting high-quality data has made it challenging for the community to develop similarly effective
19 visuo-tactile models.

20 An intuitively appealing solution is to offload data collection to robots [8, 9, 40], which can acquire
21 enormous amounts of data by repeatedly probing objects around them. However, this approach
22 captures only a narrow, “robot-centric” slice of the visuo-tactile world. The data typically is limited to
23 a specific environment (e.g., tabletops) and it fundamentally suffers from a chicken-and-egg problem,
24 as the robot must *already* be capable of touching and manipulating the objects it acquires data from.
25 In practice, this often amounts to recording data from the tabletops, typically with small objects the
26 robots can safely grasp. Recent work has also turned to simulation [22, 23], such as by modeling
27 special cases where tactile interactions can be accurately simulated (e.g., rigid objects). Yet this
28 approach, too, is highly limited. Real objects squish, deform, and bend in complex ways, and their
29 seemingly simple surfaces can hide complicated microgeometry, such as weaves of fabric and tiny
30 pores. Obtaining a full understanding of vision and touch, beyond simple robotic manipulation tasks,
31 requires modeling these subtle visuo-tactile properties.

32 We argue that many aspects of the visuo-tactile world are currently best learned by observing physical
33 interactions *performed by humans*. Humans can easily access a wide range of spaces and objects that
34 would be very challenging for robots. By capturing data from objects *in situ*, the recorded sensory
35 signals more closely match how the objects would be encountered in the wild. Inspired by this idea,
36 we present a dataset, called *Touch and Go*, in which human data collectors walk through a variety
37 of environments, probing objects with tactile sensors and simultaneously recording their actions on
38 video. Our dataset spans a wide range of indoor and outdoor environments, such as classrooms, gyms,
39 streets, and hiking trails. The objects and “stuff” [1] they contain are thus significantly more diverse
40 than those of existing datasets, making it well-suited to self-supervised learning, and to tasks that
41 require an understanding of material properties, such as visual synthesis tasks.

42 We apply our dataset to a variety of multimodal learning tasks. First, we learn tactile features through
43 self-supervised learning, by training a model to associate images with touch. We find that the learned
44 features significantly outperform supervised ImageNet [16] features on a robotic manipulation task,
45 and on recognizing materials in our dataset (Fig. 1b). Second, we propose a novel task of *tactile-driven*
46 *image stylization*: making an image “feel more like” a given tactile input. To solve this problem, we
47 adapt the recent method of Li et al. [39] to generate an image whose structure matches an input image
48 but whose style is likely to co-occur with the given tactile information. The resulting model can
49 successfully change the texture of an input image, such as by adding bumps to a smooth surface to
50 match the tactile information recorded from a rock (Fig. 1c). Finally, we study multimodal models for
51 *future touch prediction*: predicting future frames of a touch sensor’s recording, given both visual and
52 tactile signals. We show that visual information improves these predictions over touch alone (Fig. 1d).

53 **2 Related Work**

54 **Simulated vision and touch.** A variety of methods have simulated low-dimensional Biotac [56]
55 and fabric-based tactile sensing [43]. Other work has proposed to simulate high-dimensional tactile
56 data, based on visual tactile sensors such as GelSight [67, 32, 31]. Wang et al. [61] simulated visual
57 and tactile data for robotic grasps of rigid objects. Gao et al. [22, 23] proposed a dataset of simulated
58 visual, tactile, and audio data, derived from CAD models with only rigid deformation. In contrast to
59 these works, we collect our data from *real* objects and scenes, which contain non-rigid deformation,
60 microgeometry, and wider variations in visual appearance.

61 **Robotic vision and touch.** Researchers have proposed a variety of methods that use visual and
62 touch signals for robotic applications [9, 8, 14, 34, 36, 45, 41, 13, 65, 12, 19]. Several of these have

63 proposed visuo-tactile datasets. Calandra et al. created a dataset for multimodal grasping [9] and
64 regrasping [8] with a robotic arm. Li et al. [40] collected data from a robotic arm synthesis, and
65 proposed a model based on generative adversarial networks [25] for cross-modal translation. Murali
66 et al. [45] proposed a dataset for tactile-only grasping. These datasets have largely been confined to
67 specific environments (e.g., a lab space containing the robots), and only contain objects provided to
68 them by humans that the robots are capable of interacting with. Consequently, they contain a small
69 number of object instances (each less than 200).

70 **Human-collected multimodal data.** We take inspiration from work that collects data by having
71 humans physically interact with objects *in situ*. Our approach is similar to Owens et al. [48], which
72 learns audio-visual associations by probing objects with a drumstick [48]. In contrast, we collect
73 touch instead of sound, and record data in an approximately egocentric manner as they move from
74 object to object. Later work predicts the trajectory of a bounced ball [52]. Sundaram et al. [57]
75 proposed a glove that records tactile signals, and collected a dataset of human grasps for 26 objects in
76 a lab setting. Other work predicts hand pose from touch [69]. Burka et al. combined several haptic
77 sensors [5, 6]. They then demonstrated the resulting sensor by collecting a preliminary (and possibly
78 not yet released) dataset of 357 real-world surfaces and training a model to predict human ratings of
79 4 surface properties from touch. By contrast, we have significantly larger and more diverse data from
80 indoor and outdoor scenes (rather than flat, largely indoor surfaces), use a rich vision-based tactile
81 sensor (GelSight), and demonstrate our dataset on cross-modal prediction tasks.

82 **Multimodal feature learning.** In seminal work, de Sa [15] proposed to learn from correlating sight
83 from sound. A variety of methods of been proposed for training deep networks to learn features from
84 audio-visual correlations [46, 48, 49, 2, 35, 47, 44], from images and depth [59], and from vision and
85 language [33, 42, 18, 53], and matching images and touch [37]. We adapt the contrastive model of
86 Tian et al. [59] to visuo-tactile learning.

87 **Multimodal image prediction.** A variety of methods have been proposed for predicting images
88 from another modality, such as by using text or labels [55, 30, 4, 54] or sound [38, 11, 39]. Li
89 et al. [39] proposed a model for audio-driven stylization, i.e. learning to restyle images to better
90 match an audio signal. We adapt this model to tactile-driven stylization, creating a model that is
91 conditioned on tactile inputs instead of sound. We also take inspirations from work on future video
92 prediction [17, 24, 51, 62, 21, 66, 63, 3, 60]. In particular, Tian et al. [58] trains an action-conditioned
93 video prediction method to estimate future tactile signals, using a GelSight sensor controlled by a
94 CNC machine. In contrast, we predict future tactile signals from natural objects, and show that visual
95 information can improve the prediction quality.

96 3 The Touch and Go Dataset

97 We collect a dataset of natural vision-and-touch signals. Our dataset contains multimodal data
98 recorded by humans, who probe objects in their natural locations with a tactile sensor. To more easily
99 train and analyze models on this dataset, we also collect material labels and identify touch onsets.

100 3.1 Collecting a natural visuo-tactile dataset

101 To acquire our dataset, human data collectors (the authors) walked through a variety of environments,
102 probing the objects with a tactile sensor. To obtain images that show clear, zoomed-in images of
103 the objects being touched, two people collected data at once: one who presses the tactile sensor
104 onto an object, and another who records an “approximately egocentric” video (see supplement for a
105 visualization) of their hand. The two data collectors moveed from object to object in the space as
106 part of a single, continuously recording video, touching the objects around them. We show examples
107 from our dataset in Fig. 2. The captured data varies heavily in material properties (e.g., soft/hard,
108 smooth/rough), geometries, and semantics.

109 **Capturing procedure.** To ensure that our dataset captures the natural variation of real-world vision
110 and touch, we collect both rigid and deformable objects in indoor and outdoor scenes. These scenes
111 include rooms in university buildings, such as classrooms and hallways, apartments, hiking trails,
112 playgrounds, and streets. We show example footage from our model in Fig. 1, and in the supplement.
113 The data collectors selected a variety of objects in each scene to press including chairs, walls, ground,
114 sofa, table, etc. in indoor scenes, and grass, rock, tree, sand etc. in outdoor scenes, pressing each

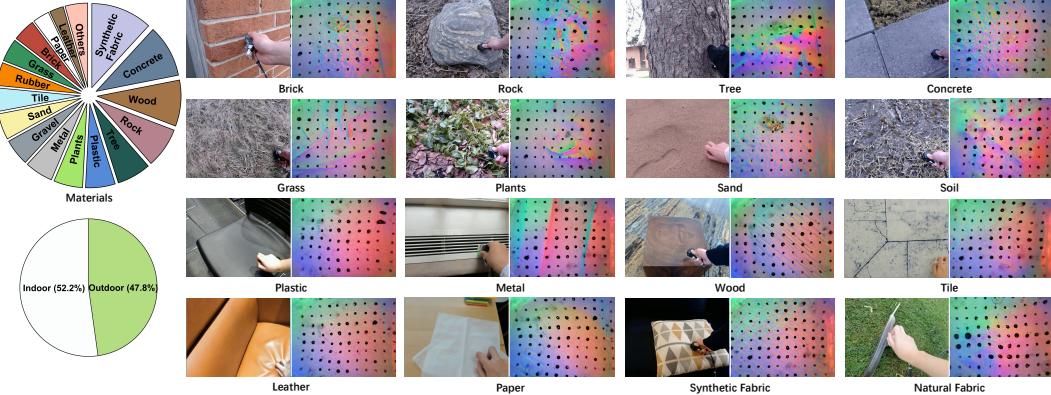


Figure 2: **The Touch and Go Dataset**. Human data collectors record paired visual and tactile information by probing objects in a variety of indoor and outdoor spaces. We show a selection of images, paired with the corresponding frame recorded by the GelSight tactile sensor. We show 16 representative categories (out of 20), and provide the distribution of material and scene types.²

one approximately 3 times. Each press lasted for 0.7 sec on average. If an object contains multiple materials (e.g., a chair with a cushion and a plastic arm), collectors generally directed their presses to each one. The data collectors also aimed to touch object parts with complex geometry, rather than flat surfaces. To avoid capturing human faces, in public spaces the captures point the camera toward the ground when moving between objects.

Hardware. For the tactile sensor, we use GelSight [32], the variant designed for robotic manipulation [67]. This is a vision-based tactile sensor, approximately 1.5cm in diameter, in which a camera observes the deformation of a curved elastomer gel illuminated by multiple colored light sources, with markers embedded inside it. The GelSight records a video in which surface orientation, depth, and shear can be estimated by analyzing the appearance of each video frame. The tactile sensor’s recordings are recorded concurrently with visual images from an ordinary webcam (both at approximately 27 Hz). Both videos (tactile and visual) are recorded on a single computer.

3.2 Annotating the dataset

To make it easier to analyze results and train models, we provide annotations for material categories and touch onsets.

Detecting touch onsets. As the data collectors move between objects, the tactile sensor does not make contact with anything. Thus, for convenience, we provide the timings of when each touch begins, so that applications can use trimmed videos (Sec. 4). To obtain these timings, we train a detector for touch onsets. We (the authors) hand-label 10k frames and train a binary ResNet-18 classifier [27] that operates on GelSight images. To ensure high accuracy, we hire workers to review the onset labels and correct errors.

Labeling materials. We label the material category for all (visual) video frames that our detector predicts as touch onsets, using a labeling scheme similar to [48]. Online workers assign a label from a list of categories. If an object is not in this category list, the workers will label it *other material*. To ensure accuracy, we have 5 workers label each image. If fewer than 4 to agree on the label, at least 2 authors watch the original video and assign a final label. We show the distribution of labels in Fig. 2.

3.3 Dataset analysis

We analyze the contents of our dataset and compare it to other works.

Data distribution. In Fig. 2, we show statistics of labeled materials and scene types, and provide qualitative results from the dataset. It contains approximately 13.9k detected touches and approximately 3971 individual object instances. Since we do not explicitly label instances, we obtain the latter number by exhaustively counting the objects in 10% of the videos and extrapolating. Our dataset is relatively balanced between indoor (52.2%) and outdoor (47.8%) scenes. We found that

²Following common practice, the tactile images are enhanced for visualization purpose. Contrast and sharpness are increased by 30% and saturation is increased by 20%.

Table 1: **Tactile datasets.** We compare attributes of our dataset with several previously proposed datasets.

| | Object inst. | Touches | Source | Real-world | Environment |
|----------------------------|--------------|--------------|--------------|------------|-----------------------|
| More Than a Feeling [8] | 65 | 6.5k | Robot | ✓ | Tabletop |
| The Feeling of Success [9] | 106 | 9.3k | Robot | ✓ | Tabletop |
| VisGel [40] | 195 | 12k | Robot | ✓ | Tabletop |
| ObjectFolder 1.0 [22] | 100 | - | Synthetic | ✗ | - |
| ObjectFolder 2.0 [23] | 1000 | - | Synthetic | ✗ | - |
| Burka et al. [7] | 357 | 1.1k | Human | ✓ | Mostly Indoor |
| Touch and Go (Ours) | 3971 | 13.9k | Human | ✓ | Indoor/Outdoor |

several categories, namely *synthetic fabric*, *tile*, *paper*, and *leather*, are only present in our indoor scenes, while *tree*, *grass*, *plant*, and *sand* are only present in outdoor scenes. The remaining materials exist in both scenes. We provide more details in the supplement.

Comparison to other datasets. In Table 1, we compare our dataset to several previously proposed visuo-tactile datasets collected by robots, by humans, or through simulation. Our dataset contains approximately $4\times$ as many object instances as the second-largest dataset, the simulation-based ObjectFolder 2.0 [23], and $11\times$ larger compared with the human-collected dataset by Burka et al. [7]. Compared to the existing robot-collected datasets, ours contains more touches (e.g., $1.15\times$ more than VisGel [40] and $1.5\times$ more than The Feeling of Success [9]). Our dataset also contains data from more diverse scenes than prior work, with a mixture of natural indoor and outdoor scenes. In contrast, robot-collected datasets [8, 9, 40] are confined to a single lab space containing the robot.

4 Applications

To evaluate the effectiveness of our dataset, we perform tasks that are designed to span a variety of application domains, including representation learning, image synthesis, and future prediction.

4.1 Multimodal self-supervised representation learning

We ask, first, whether we can use the multimodal data to learn representations for the tactile modality by associating touch with sight. We then ask how well the learned representations convey material properties from our dataset, and whether they are useful for robotic learning tasks whose data has been collected in other works [8, 9]. The latter task requires a significant amount of generalization, since the objects manipulated by robots while our dataset is collected by humans.

Our goal is to learn a representation that captures the information and correspondences between visual and tactile images, which can be useful for downstream tasks. Given the visual and tactile datasets, X_I and X_T , we aim to extract the corresponding visual-tactile pairs, $\{\mathbf{x}_I^i, \mathbf{x}_T^i\}$ and mismatched pairs $\{\mathbf{x}_I^i, \mathbf{x}_T^j\}$ using the Contrastive Multiview Coding (CMC) model proposed by Tian et al. [59]. The detailed procedure is shown below.

For each visual-tactile image pair, we first encode visual images and tactile inputs as L2-normalized embeddings using two networks, where $z_I = f_{\theta_I}(\mathbf{x}_I)$ for visual images and $z_T = f_{\theta_T}(\mathbf{x}_T)$ for tactile images. Recall that our goal is to find the corresponding sample of the other modality, given a set S that contains both the corresponding example and $K - 1$ random examples. When matching a visual example \mathbf{x}_I^i to tactile, the loss for matching visual images to tactile images is:

$$\mathcal{L}_{\text{contrast}}^{X_I, X_T} = -\log \frac{\exp(f_{\theta_I}(\mathbf{x}_I^i) \cdot f_{\theta_T}(\mathbf{x}_T^i)/\tau)}{\sum_{j=1}^K \exp(f_{\theta_I}(\mathbf{x}_I^i) \cdot f_{\theta_T}(\mathbf{x}_T^j)/\tau)} \quad (1)$$

where $\tau = 0.07$ is a constant and j indexes the tactile examples in S . Analogously, we can obtain a loss in which tactile examples are matched to images, $\mathcal{L}_{\text{contrast}}^{X_T, X_I}$. We minimize both losses:

$$\mathcal{L}(X_I, X_T) = \mathcal{L}_{\text{contrast}}^{X_I, X_T} + \mathcal{L}_{\text{contrast}}^{X_T, X_I} \quad (2)$$

The training details and hyperparameters are provided in the supplementary material.

181 **4.2 Tactile-driven image stylization**

182 Both touch and sight convey material properties and geometry. A model that can successfully predict
183 these properties from visuo-tactile data therefore ought to be able to translate between modalities.
184 Inspired by the audio-driven image stylization of Li et al. [39], we propose the task of *tactile-driven*
185 *image stylization*: making an image look as though it “feels like” a given touch signal.

186 Following Li et al. [39], we base our approach on contrastive unpaired translation (CUT) [50]. Given
187 an input image \mathbf{x}_I and a tactile example \mathbf{x}_T , our goal is to manipulate the image via a generator
188 such that the manipulated pair $\hat{\mathbf{x}}_I = G(\mathbf{x}_I, \mathbf{x}_T)$ is more likely to co-occur in the dataset. Our model
189 consists of an image translation network that is conditioned on a tactile example (a GelSight image).
190 The loss function encourages the model to preserve the image structure, which is enforced using an
191 image-based contrastive loss [50], while adjusting the image style such that the resulting textures are
192 more likely to co-occur with the tactile signal, which is measured using a visuo-tactile discriminator.
193 **Making sight consistent with touch.** We train the model with a discriminator D that distinguishes
194 between real (and fake) visuo-tactile pairs. During training time, we shuffle the dataset to generate the
195 set S_n containing mismatched image-tactile pairs $\{\mathbf{x}_I, \mathbf{x}'_T\} \in S_n$. Likewise, we define the original
196 dataset as S_m which contains matched image-tactile pairs $\{\mathbf{x}_I, \mathbf{x}_T\} \in S_m$. In formal terms, the
197 visual-tactile adversarial loss can be written as:

$$\mathcal{L}_{\text{VT}} = \mathbb{E}_{\{\mathbf{x}_I, \mathbf{x}_T\} \sim S_m} \log D(\mathbf{x}_I, \mathbf{x}_T) + \mathbb{E}_{\{\mathbf{x}_I, \mathbf{x}'_T\} \sim S_n} \log(1 - D(G(\mathbf{x}_I, \mathbf{x}'_T), \mathbf{x}'_T)) \quad (3)$$

198 **Loss.** We combine our visuo-tactile discriminator with the structure-preserving loss used in Li
199 et al. [40], which was originally proposed by CUT [50]. This loss, which we call \mathcal{L}_{CUT} , works
200 by training a contrastive learning model that puts patches in the input and predicted images into
201 correspondence, such that patches at the same positions are close in an embedding space. Please see
202 the supplement for more details. The overall loss is:
203

$$\mathcal{L}_{\text{TDIS}} = \mathcal{L}_{\text{VT}} + \mathcal{L}_{\text{CUT}}. \quad (4)$$

204 **4.3 Multimodal future touch prediction**

205 Inspired by the challenges of action-conditional future touch prediction [58], we use our dataset to
206 ask whether *visual* data can improve our estimates of future tactile signals: i.e., what will this object
207 feel like in a moment? Visual information can help touch prediction in a number of ways, such as by
208 conveying material properties (e.g., deformation) and geometry. It can also provide information
209 about which action is being performed. One may thus also consider it as an *implicit* form of action
210 conditioning [20], since the visual information provides information analogous to actions.

211 We adapt the video prediction architecture of Geng et al. [24] to this task. This model is based
212 on residual networks [27, 29] with 3D space-time convolutions (please see the supplement for
213 architecture details). We predict multiple frames by autoregressively feeding our output images
214 back to the original model. Given a series of paired visual and tactile images from times 1 to t ,
215 $\{(\mathbf{x}_I^1, \mathbf{x}_T^1), \dots, (\mathbf{x}_I^t, \mathbf{x}_T^t)\}$, our goal is to predict the subsequent tactile image, $\hat{\mathbf{x}}_T^{(t+1)}$. We train the
216 model with L1 and perceptual loss, following [24].

217 **5 Experiments**

218 **5.1 Self-supervised feature learning**

219 We train a self-supervised model that learns to associate images with touch. We evaluate this learned
220 representation on two downstream tasks: robot grasping and material prediction.

221 **Training data.** Since the dataset was concurrently collected with the applications, for this and all
222 applications, we use the initial 10 hours of data (approximately 10k touches), following common
223 practice in other dataset papers [10].

224 **Robotic grasping task.** For the robot grasping task, we use the experimental setup and dataset
225 of Calandra et al. [9]. Thus, the task requires generalizing to data recorded in a very different
226 environment, and with different GelSight sensors. The goal of this task is to predict whether a
227 robotic arm will successfully grasp an object, based on inputs from two GelSight images recorded

Table 2: Linear probing on robot grasping and material classification tasks. We provide more results on other tactile datasets in supplementary materials.

| Dataset | Method | Robot Grasping Accuracy(%) | | Material Classification Accuracy(%) | |
|---------------|------------------------|----------------------------|-------------|-------------------------------------|-------------|
| | | ResNet-18 | ResNet-50 | ResNet-18 | ResNet-50 |
| - | Chance | 56.1 | 56.1 | 18.6 | 18.6 |
| ImageNet [16] | Supervised | 73.0 | 75.6 | 31.2 | 33.4 |
| ImageNet [16] | MoCo [26] | - | 75.9 | - | 34.1 |
| ImageNet [16] | L/ab CMC [59] | - | 76.9 | - | 31.3 |
| Ours | Visuo-tactile CMC [59] | 78.6 | 81.5 | 35.5 | 36.3 |

before and after grasping. Since there is no standard training/test split from [9], we split their objects randomly into training/test. The resulting dataset contains 68 objects and 5921 touches for training, 16 objects and 1204 touches for validation, and 21 objects and 1204 touches for testing. Similar to the tactile-only model from Calandra et al. [9], we compute features for each of the 4 tactile images (before/after images for 2 tactile sensors) using our self-supervised model. We concatenate these features together and train a linear classifier to solve this binary classification task.

Material classification task. For the material classification task, we use our own dataset. The goal of this task is to predict the material label from a tactile image. To avoid providing our self-supervised learning model with object instances that appear in the linear probing experiment, we split the dataset into an unlabeled set containing 5172 touches (51.7%), a labeled training set of 3921 touches (39.2%), and a labeled test set of 923 touches (9.1%). We split the dataset by video (rather than by touch or frame) to avoid having the same (or nearly the same) object appear in both training and test.

Implementation details. We train our model for 240 epochs, using the optimization parameters from CMC [59], after adjusting the learning rate schedule to compensate for longer training. We set the weight decay to be 10^{-4} . We train our model with the batch size of 128 on 4 Nvidia 2080-Ti GPUs. For the downstream classification tasks, we froze our network weights and obtained visual features by performing global average pooling on the final convolutional layer. We follow the approach of [59] for learning the linear classifier.

Comparison to other feature sets. We show downstream classification results in Table 2. To evaluate the effectiveness of our dataset, we compare our learned features to those of several other approaches. These include using supervised ImageNet [16] features, which are commonly used to represent GelSight images [68, 9, 8], and visual CMC [59] features trained on ImageNet, which treats the L and ab color channels of an image as different modalities. We see that our model obtains significantly better performance than these models on both tasks, due to its ability to learn from real tactile data. These results suggest that our dataset provides a useful signal for training self-supervised tactile representations.

5.2 Tactile-driven image stylization

We use our model to modify the style of an image to match a touch signal.

Implementation details. Following Li et al. [39], during training we sample a random image from the dataset, along with a second visuo-tactile pair, and use the pair to restyle the image, using the loss in Eq. 4. We provide architectural details in the supplement. For the discriminator we adopt the PatchGAN architecture [28]. The architecture of discriminator follows [39], which concatenates the two input images channel-wise, and passes the combined images to the discriminator. We train our model on 4 Nvidia 2080-Ti GPUs for 100 epochs with a batch size of 8 and the learning rate of 0.0002. We augment the vision images with random cropping and horizontal flipping.

Experimental setup. Following [39], we evaluate our model by restyling images in our dataset with random tactile inputs, both of which are taken from the test set. We use evaluation metrics that measure the consistency of the manipulated image with the example used for conditioning. First, we measure the similarity of the manipulated image and the tactile example used for conditioning. Similar to [39], we use our trained CMC model, by taking the dot product between visual and tactile embeddings. Second, we compare

Table 3: Quantitative results for tactile-driven image stylization.

| Method | CMC | Material |
|---------------|--------------|--------------|
| Baseline | 0.165 | 0.107 |
| CycleGAN [71] | 0.178 | 0.129 |
| Ours | 0.197 | 0.142 |



Figure 3: Qualitative results our model on tactile-driven image stylization. For each row, we show an input image (left) and the manipulated image (to its right) obtained by stylizing with a given tactile input (right side). For reference, we also show the image that corresponds to the tactile example at rightmost (not used by the model). The manipulated images convey physical properties of the tactile signal, such as its roughness (e.g., first three rows) or smoothness (e.g., row 10). Other inputs result in images that combine the properties of two inputs (e.g., by adding grass, as in row 8). We also show failure cases in the last row. *Zoom in for better view.*

material prediction consistency between the manipulated image with the (held out) conditional image. We use our material classifier to categorize the predicted and conditioning images, and measure the rate at which they agree. Since this is a novel task, we create a second variation of our model for comparison, following [71]. This model performs the stylization using CycleGAN, rather than CUT (see supplement for details).

Quantitative results. Quantitative results are shown in Table 3. Here, the “baseline” indicates results from original image before stylization. We can see that the CUT-based method obtains higher CMC similarity than a CycleGAN-based method [71]. In terms of material classification consistency, our model consistently outperforms CycleGAN-based method [71].

Qualitative results. In Fig. 3, we show results from our model. Our model successfully manipulates images to match tactile inputs, such as by making surfaces rougher or smoother, or by creating “hybrid” materials (e.g., adding grass to a surface). These results are obtained *without* having access to the tactile example’s corresponding image, suggesting that the model has learned which physical properties are shared between sight and touch.

Table 4: Quantitative results for video prediction

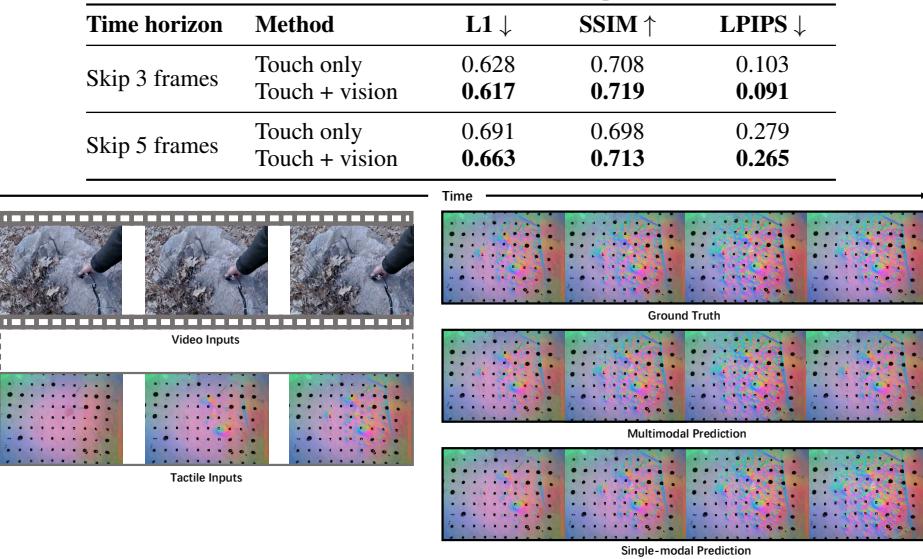


Figure 4: Future touch prediction. We show the results of tactile-only and visuo-tactile models.

285

5.3 Multimodal video prediction

286 We evaluate our model for predicting future tactile signals. We compare a tactile-only model to a
 287 multimodal visuo-tactile model, and show that the latter obtains better performance.

288 **Experimental setting.** We evaluate the effectiveness of multimodal inputs using three context
 289 frames to predict the next frame under two different time horizons: skipping 3 and 5 frames between
 290 contexts. Following [24], we adopt three evaluation metrics: MAE, SSIM [64] and LPIPS [70]. We
 291 provide training hyperparameters in the supplement.

292 **Multimodal vs. single-modal prediction.** We show the quantitative results in Table 4. We can see
 293 that, by incorporating our dataset’s visual signal, the model gains a constant performance increase
 294 under different evaluation metrics, under both experimental settings. The gap becomes larger for
 295 longer time horizon, suggesting that visual information may be more helpful in this case.

296

6 Discussion

297 We proposed *Touch and Go*, a human-collected visuo-tactile dataset. Our dataset comes from real-
 298 world objects, and is significantly more diverse than prior datasets. We demonstrate its effectiveness
 299 on a variety of applications that involve robotic manipulation, material understanding, and image
 300 synthesis. In the tradition of previous work [48], we see our work as a step toward *human-collected*
 301 multimodal data collection, in which humans equipped with multiple sensors collect diverse dataset by
 302 recording themselves physically interacting with the world. We hope this data will enable researchers
 303 to study diverse visuo-tactile learning applications, beyond the “robotics-centric” domains that are
 304 often the focus of previous efforts.

305 **Limitations.** Collecting diverse tactile data is an ongoing challenge, since it requires physically
 306 being present in the locations where data is collected. While adding human collectors improves
 307 diversity in many ways, our dataset was mainly collected in one geographic location (near University
 308 of Michigan’s campus). Consequently, the data we recorded may not generalize to all spaces. The use
 309 of humans in the data collection process also potentially introduces bias, which differs from “robotic”
 310 or “virtual data” bias. For example, humans may choose unrepresentative parts of the objects to
 311 probe, and do not perform actions with consistent force. The humans who recorded the dataset may
 312 also not be representative of the general population, which may introduce bias (e.g., in skin tone). To
 313 help explain these biases, we provide more detailed information about the dataset in the supplement.

314 **Acknowledgements.** We thank Xiaofeng Guo and Yufan Zhang for the extensive help with the
 315 GelSight sensor, and thank Daniel Geng, Yuexi Du and Zhaoying Pan for the helpful discussions.
 316 This work was supported in part by Cisco Systems.

317 **References**

- 318 [1] Edward H Adelson. On seeing stuff: the perception of materials by humans and machines. In *Human
319 vision and electronic imaging VI*, volume 4299, pages 1–12. SPIE, 2001.
- 320 [2] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International
321 Conference on Computer Vision*, pages 609–617, 2017.
- 322 [3] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H Campbell, and Sergey Levine. Stochastic
323 variational video prediction. *arXiv preprint arXiv:1710.11252*, 2017.
- 324 [4] David Bau, Alex Andonian, Audrey Cui, YeonHwan Park, Ali Jahanian, Aude Oliva, and Antonio Torralba.
325 Paint by word. In *arXiv:2103.10951*, 2021.
- 326 [5] Alex Burka, Siyao Hu, Stuart Helgeson, Shweta Krishnan, Yang Gao, Lisa Anne Hendricks, Trevor Darrell,
327 and Katherine J Kuchenbecker. Proton: A visuo-haptic data acquisition system for robotic learning of
328 surface properties. In *2016 IEEE International Conference on Multisensor Fusion and Integration for
329 Intelligent Systems (MFI)*, pages 58–65. IEEE, 2016.
- 330 [6] Alex Burka, Abhinav Rajvanshi, Sarah Allen, and Katherine J Kuchenbecker. Proton 2: Increasing the
331 sensitivity and portability of a visuo-haptic surface interaction recorder. In *2017 IEEE International
332 Conference on Robotics and Automation (ICRA)*, pages 439–445. IEEE, 2017.
- 333 [7] Alexander L Burka. Instrumentation, data, and algorithms for visually understanding haptic surface
334 properties. *Dissertations available from ProQuest*, 2018.
- 335 [8] Roberto Calandra, Andrew Owens, Dinesh Jayaraman, Justin Lin, Wenzhen Yuan, Jitendra Malik, Ed-
336 ward H. Adelson, and Sergey Levine. More than a feeling: Learning to grasp and regrasp using vision and
337 touch. *IEEE Robotics and Automation Letters*, 3:3300–3307, 2018.
- 338 [9] Roberto Calandra, Andrew Owens, Manu Upadhyaya, Wenzhen Yuan, Justin Lin, Edward H Adelson, and
339 Sergey Levine. The feeling of success: Does touch sensing help predict grasp outcomes? *Conference on
340 Robot Learning (CoRL)*, 2017.
- 341 [10] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset.
342 In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308,
343 2017.
- 344 [11] Moitrey Chatterjee and Anoop Cherian. Sound2sight: Generating visual dynamics from sound and
345 context. In *European Conference on Computer Vision*, pages 701–719. Springer, 2020.
- 346 [12] Vivian Chu, Ian McMahon, Lorenzo Riano, Craig G McDonald, Qin He, Jorge Martinez Perez-Tejada,
347 Michael Arrigo, Trevor Darrell, and Katherine J Kuchenbecker. Robotic learning of haptic adjectives
348 through physical interaction. *Robotics and Autonomous Systems*, 63:279–292, 2015.
- 349 [13] Shaowei Cui, Rui Wang, Junhang Wei, Jingyi Hu, and Shuo Wang. Self-attention based visual-tactile
350 fusion learning for predicting grasp outcomes. *IEEE Robotics and Automation Letters*, 5(4):5827–5834,
351 2020.
- 352 [14] Mark R. Cutkosky, Robert D. Howe, and William R. Provancher. *Force and Tactile Sensors*, pages 455–476.
353 Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.
- 354 [15] Virginia R de Sa. Learning classification with unlabeled data. *Advances in neural information processing
355 systems*, 1994.
- 356 [16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical
357 image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255,
358 2009.
- 359 [17] Emily L. Denton and Rob Fergus. Stochastic video generation with a learned prior. In *ICML*, 2018.
- 360 [18] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In
361 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11162–
362 11173, 2021.
- 363 [19] Nima Fazeli, Miquel Oller, Jiajun Wu, Zheng Wu, Joshua B Tenenbaum, and Alberto Rodriguez. See, feel,
364 act: Hierarchical learning for complex manipulation skills with multisensory fusion. *Science Robotics*,
365 4(26):eaav3123, 2019.

- 366 [20] Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through
 367 video prediction. *Advances in neural information processing systems*, 29, 2016.
- 368 [21] Hang Gao, Huazhe Xu, Qi-Zhi Cai, Ruth Wang, Fisher Yu, and Trevor Darrell. Disentangling propagation
 369 and generation for video prediction. In *Proceedings of the IEEE/CVF International Conference on*
 370 *Computer Vision*, pages 9006–9015, 2019.
- 371 [22] Ruohan Gao, Yen-Yu Chang, Shivani Mall, Li Fei-Fei, and Jiajun Wu. Objectfolder: A dataset of objects
 372 with implicit visual, auditory, and tactile representations. In *CoRL*, 2021.
- 373 [23] Ruohan Gao, Zilin Si, Yen-Yu Chang, Samuel Clarke, Jeannette Bohg, Li Fei-Fei, Wenzhen Yuan, and
 374 Jiajun Wu. Objectfolder 2.0: A multisensory object dataset for sim2real transfer. In *CVPR*, 2022.
- 375 [24] Daniel Geng, Max Hamilton, and Andrew Owens. Comparing correspondences: Video prediction with
 376 correspondence-wise losses. *Computer Vision and Pattern Recognition (CVPR)*, 2022.
- 377 [25] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron
 378 Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes,
 379 N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.
- 381 [26] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised
 382 visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and*
 383 *Pattern Recognition (CVPR)*, June 2020.
- 384 [27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition.
 385 *arXiv preprint arXiv:1512.03385*, 2015.
- 386 [28] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional
 387 adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*,
 388 pages 5967–5976, 2017.
- 389 [29] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and
 390 super-resolution, 2016.
- 391 [30] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the*
 392 *IEEE conference on computer vision and pattern recognition*, pages 1219–1228, 2018.
- 393 [31] Micah Johnson, Forrester Cole, Alvin Raj, and Edward Adelson. Microgeometry capture using an
 394 elastomeric sensor. *ACM Trans. Graph.*, 30:46, 07 2011.
- 395 [32] Micah K Johnson and Edward H Adelson. Retrographic sensing for the measurement of surface texture
 396 and shape. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1070–1077.
 397 IEEE, 2009.
- 398 [33] Armand Joulin, Laurens van der Maaten, Allan Jabri, and Nicolas Vasilache. Learning visual features from
 399 large weakly supervised data. In *European Conference on Computer Vision*, pages 67–84. Springer, 2016.
- 400 [34] Zhanat Kappassov, Juan-Antonio Corrales, and Véronique Perdereau. Tactile sensing in dexterous robot
 401 hands — review. *Robotics and Autonomous Systems*, 74:195–220, 2015.
- 402 [35] Bruno Korbar, Du Tran, and Lorenzo Torresani. Co-training of audio and video representations from
 403 self-supervised temporal synchronization. *CoRR*, 2018.
- 404 [36] Susan J Lederman and Roberta L Klatzky. Hand movements: A window into haptic object recognition.
 405 *Cognitive Psychology*, 19(3):342–368, 1987.
- 406 [37] Michelle A Lee, Yuke Zhu, Krishnan Srinivasan, Parth Shah, Silvio Savarese, Li Fei-Fei, Animesh
 407 Garg, and Jeannette Bohg. Making sense of vision and touch: Self-supervised learning of multimodal
 408 representations for contact-rich tasks. *arXiv preprint arXiv:1810.10191*, 2018.
- 409 [38] Seung Hyun Lee, Wonseok Roh, Wonmin Byeon, Sang Ho Yoon, Chan Young Kim, Jinkyu Kim, and
 410 Sangpil Kim. Sound-guided semantic image manipulation. *arXiv preprint arXiv:2112.00007*, 2021.
- 411 [39] Tingle Li, Yichen Liu, Andrew Owens, and Hang Zhao. Learning visual styles from audio-visual
 412 associations. *arXiv*, 2022.
- 413 [40] Yunzhu Li, Jun-Yan Zhu, Russ Tedrake, and Antonio Torralba. Connecting touch and vision via cross-
 414 modal prediction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June
 415 2019.

- 416 [41] Justin Lin, Roberto Calandra, and Sergey Levine. Learning to identify object instances by touch: Tactile
 417 recognition via multimodal matching. In *2019 International Conference on Robotics and Automation*
 418 (*ICRA*), pages 3644–3650. IEEE, 2019.
- 419 [42] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin
 420 Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In
 421 *Proceedings of the European conference on computer vision (ECCV)*, pages 181–196, 2018.
- 422 [43] Andrew Melnik, Luca Lach, Matthias Plappert, Timo Korthals, Robert Haschke, and Helge Ritter. Tactile
 423 sensing and deep reinforcement learning for in-hand manipulation tasks. In *IROS Workshop on Autonomous*
 424 *Object Manipulation*, 2019.
- 425 [44] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-
 426 modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni-*
 427 *tion*, pages 12475–12486, 2021.
- 428 [45] Adithyavairavan Murali, Yin Li, Dhiraj Gandhi, and Abhinav Gupta. Learning to grasp without seeing. In
 429 *International Symposium on Experimental Robotics*, pages 375–386. Springer, 2018.
- 430 [46] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal
 431 deep learning. In *ICML*, 2011.
- 432 [47] Andrew Owens and Alexei A. Efros. Audio-visual scene analysis with self-supervised multisensory
 433 features. *Eur. Conf. Comput. Vis.*, 2018.
- 434 [48] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H. Adelson, and William T.
 435 Freeman. Visually indicated sounds. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*
 436 (*CVPR*), pages 2405–2413, 2016.
- 437 [49] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient
 438 sound provides supervision for visual learning. In *Eur. Conf. Comput. Vis.*, 2016.
- 439 [50] Taesung Park, Alexei A. Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for conditional
 440 image synthesis. In *ECCV*, 2020.
- 441 [51] Viorica Patrascu, Ankur Handa, and Roberto Cipolla. Spatio-temporal video autoencoder with differen-
 442 tiable memory. *arXiv preprint arXiv:1511.06309*, 2015.
- 443 [52] Senthil Purushwalkam, Abhinav Gupta, Danny M Kaufman, and Bryan Russell. Bounce and learn:
 444 Modeling scene dynamics with real-world bounces. *arXiv preprint arXiv:1904.06827*, 2019.
- 445 [53] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish
 446 Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from
 447 natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR,
 448 2021.
- 449 [54] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and
 450 Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021.
- 451 [55] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee.
 452 Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, pages
 453 1060–1069, 2016.
- 454 [56] Philipp Ruppel, Yannick Jonetzko, Michael Görner, Norman Hendrich, and Jianwei Zhang. Simulation
 455 of the syntouch biotac sensor. In *International Conference on Intelligent Autonomous Systems*, pages
 456 374–387. Springer, 2018.
- 457 [57] Subramanian Sundaram, Petr Kellnhofer, Yunzhu Li, Jun-Yan Zhu, Antonio Torralba, and Wojciech
 458 Matusik. Learning the signatures of the human grasp using a scalable tactile glove. *Nature*, 569(7758):698–
 459 702, 2019.
- 460 [58] Stephen Tian, Frederik Ebert, Dinesh Jayaraman, Mayur Mudigonda, Chelsea Finn, Roberto Calandra,
 461 and Sergey Levine. Manipulation by feel: Touch-based control with deep predictive models. In *2019*
 462 *International Conference on Robotics and Automation (ICRA)*, pages 818–824. IEEE, 2019.
- 463 [59] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference*
 464 *on computer vision*, pages 776–794. Springer, 2020.

- 465 [60] Ruben Villegas, Arkanath Pathak, Harini Kannan, Dumitru Erhan, Quoc V. Le, and Honglak Lee. High
466 fidelity video prediction with large stochastic recurrent neural networks, 2019.
- 467 [61] Shaoxiong Wang, Mike Lambeta, Po-Wei Chou, and Roberto Calandra. Tacto: A fast, flexible, and
468 open-source simulator for high-resolution vision-based tactile sensors. *IEEE Robotics and Automation*
469 *Letters*, 7(2):3930–3937, 2022.
- 470 [62] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. Few-shot
471 video-to-video synthesis. *arXiv preprint arXiv:1910.12713*, 2019.
- 472 [63] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-
473 resolution image synthesis and semantic manipulation with conditional gans, 2017.
- 474 [64] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error
475 visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.
- 476 [65] Youngsun Wi, Pete Florence, Andy Zeng, and Nima Fazeli. Virdo: Visio-tactile implicit representations of
477 deformable objects. *arXiv preprint arXiv:2202.00868*, 2022.
- 478 [66] Yue Wu, Rongrong Gao, Jaesik Park, and Qifeng Chen. Future video synthesis with object motion
479 prediction. In *CVPR*, 2020.
- 480 [67] Wenzhen Yuan, Siyuan Dong, and Edward H Adelson. Gelsight: High-resolution robot tactile sensors for
481 estimating geometry and force. *Sensors*, 17(12):2762, 2017.
- 482 [68] Wenzhen Yuan, Chenzhuo Zhu, Andrew Owens, Mandayam A Srinivasan, and Edward H Adelson.
483 Shape-independent hardness estimation using deep learning and a gelsight tactile sensor. In *International*
484 *Conference on Robotics and Automation (ICRA)*, 2017.
- 485 [69] Qiang Zhang, Yunzhu Li, Yiyue Luo, Wan Shou, Michael Foshey, Junchi Yan, Joshua B Tenenbaum,
486 Wojciech Matusik, and Antonio Torralba. Dynamic modeling of hand-object interactions via tactile sensing.
487 In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2874–2881.
488 IEEE, 2021.
- 489 [70] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable
490 effectiveness of deep features as a perceptual metric, 2018.
- 491 [71] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using
492 cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference*
493 *on*, 2017.

494 **Checklist**

- 495 1. For all authors...
- 496 (a) Do the main claims made in the abstract and introduction accurately reflect the paper's
497 contributions and scope? **[Yes]**
- 498 (b) Did you describe the limitations of your work? **[Yes]**
- 499 (c) Did you discuss any potential negative societal impacts of your work? **[N/A]** No
500 potential negative social impacts.
- 501 (d) Have you read the ethics review guidelines and ensured that your paper conforms to
502 them? **[Yes]**
- 503 2. If you are including theoretical results...
- 504 (a) Did you state the full set of assumptions of all theoretical results? **[N/A]** No theoretical
505 results
- 506 (b) Did you include complete proofs of all theoretical results? **[N/A]** No theoretical results
- 507 3. If you ran experiments (e.g. for benchmarks)...
- 508 (a) Did you include the code, data, and instructions needed to reproduce the main experi-
509 mental results (either in the supplemental material or as a URL)? **[Yes]**
- 510 (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they
511 were chosen)? **[Yes]**

- (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? **[No]**

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? **[Yes]**

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

 - (a) If your work uses existing assets, did you cite the creators? **[Yes]**
 - (b) Did you mention the license of the assets? **[No]**
 - (c) Did you include any new assets either in the supplemental material or as a URL? **[Yes]**
 - (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? **[Yes]**
 - (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **[Yes]**

5. If you used crowdsourcing or conducted research with human subjects...

 - (a) Did you include the full text of instructions given to participants and screenshots, if applicable? **[N/A]** No human subject exist in our dataset.
 - (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? **[N/A]** No human subjects exist in our dataset.
 - (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? **[N/A]** No human subjects exist in our dataset.