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

Fengyu Yang[1][*]   **Chenyang Ma**[1][*]   **Jiacheng Zhang**[1]
Jing Zhu[1]   **Wenzhen Yuan**[2]   **Andrew Owens**[1]

[1]University of Michigan   [2]Carnegie Mellon University
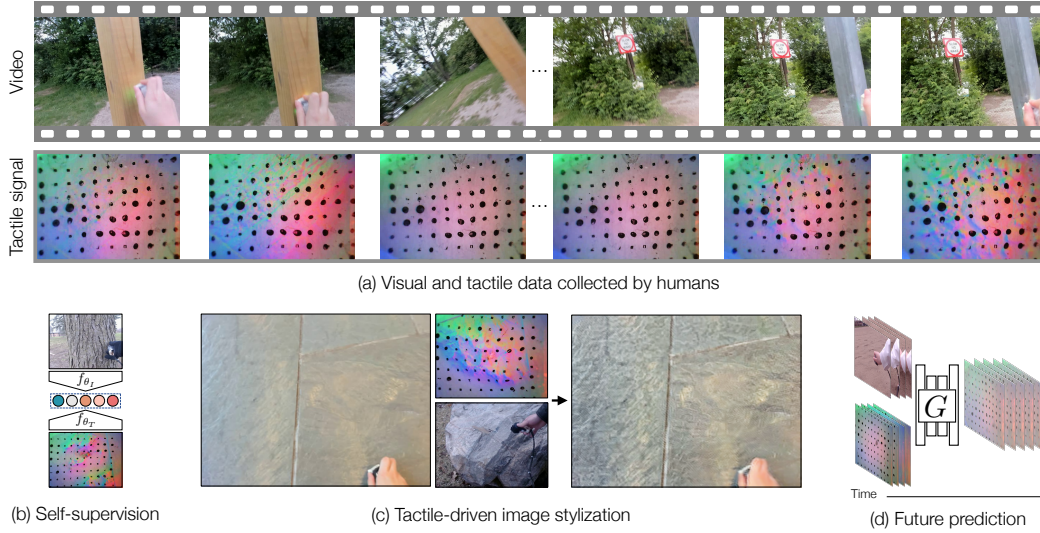
https://touch-and-go.github.io

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 tactile signals from visuo-tactile inputs.

Abstract

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

---

[*]The first two authors contributed equally to this paper.

1 Introduction

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

An intuitively appealing solution is to offload data collection to robots [8, 9, 43], which can acquire enormous amounts of data by repeatedly probing objects around them. However, this approach captures only a narrow, "robot-centric" slice of the visuo-tactile world. The data typically is limited to a specific environment (e.g., tabletops) and it fundamentally suffers from a chicken-and-egg problem, as the robot must *already* be capable of touching and manipulating the objects it acquires data from. In practice, this often amounts to recording data from the tabletops, typically with small objects the robots can safely grasp. Recent work has also turned to simulation [22, 23], such as by modeling special cases where tactile interactions can be accurately simulated (e.g., rigid objects). Yet this approach, too, is highly limited. Real objects squish, deform, and bend in complex ways, and their seemingly simple surfaces can hide complicated microgeometry, such as weaves of fabric and tiny pores. Obtaining a full understanding of vision and touch, beyond simple robotic manipulation tasks, requires modeling these subtle visuo-tactile properties.

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

We apply our dataset to a variety of multimodal learning tasks. First, we learn tactile features through self-supervised learning, by training a model to associate images with touch. We find that the learned features significantly outperform supervised ImageNet [16] features on a robotic manipulation task, and on recognizing materials in our dataset (Fig. 1b). Second, we propose a novel task of *tactile-driven image stylization*: making an image "feel more like" a given tactile input. To solve this problem, we adapt the recent method of Li et al. [42] to generate an image whose structure matches an input image but whose style is likely to co-occur with the given tactile information. This task evaluates the ability to learn cross-modal associations – i.e., how an object feels from how it looks and vice versa. The resulting model can successfully change the texture of an input image, such as by adding bumps to a smooth surface to match the tactile information recorded from a rock (Fig. 1c). Finally, we study multimodal models for *future touch prediction*: predicting future frames of a touch sensor's recording, given both visual and tactile signals. We show that visual information improves these predictions over touch alone (Fig. 1d).

2 Related Work

Simulated vision and touch. A variety of methods have simulated low-dimensional Biotac [59] and fabric-based tactile sensing [46]. Other work has proposed to simulate high-dimensional tactile data, based on visual tactile sensors such as GelSight [71, 33, 32]. Wang et al. [64] simulated visual and tactile data for robotic grasps of rigid objects. Gao et al. [22, 23] proposed a dataset of simulated visual, tactile, and audio data, derived from CAD models with only rigid deformation. In contrast to these works, we collect our data from *real* objects and scenes, which contain non-rigid deformation, microgeometry, and wider variations in visual appearance.

Robotic vision and touch. Researchers have proposed a variety of methods that use visual and touch signals for robotic applications [9, 8, 14, 35, 38, 48, 44, 13, 68, 12, 19]. Several of these have proposed visuo-tactile datasets. Calandra et al. created a dataset for multimodal grasping [9] and regrasping [8] with a robotic arm. Li et al. [43] collected data from a robotic arm synthesis, and

proposed a model based on generative adversarial networks [25] for cross-modal translation. Murali et al. [48] proposed a dataset for tactile-only grasping. These datasets have largely been confined to specific environments (e.g., a lab space containing the robots), and only contain objects provided to them by humans that the robots are capable of interacting with. Consequently, they contain a small number of object instances (each less than 200).

Human-collected multimodal data. We take inspiration from work that collects data by having humans physically interact with objects *in situ*. Our approach is similar to Owens et al. [51], which learns audio-visual associations by probing objects with a drumstick [51]. In contrast, we collect touch instead of sound, and record data in an approximately egocentric manner as they move from object to object. Later work predicts the trajectory of a bounced ball [55]. Sundaram et al. [60] proposed a glove that records tactile signals, and collected a dataset of human grasps for 26 objects in a lab setting. Other work predicts hand pose from touch [73]. Burka et al. combined several haptic sensors [5, 6]. They then demonstrated the resulting sensor by collecting a preliminary (and possibly not yet released) dataset of 357 real-world surfaces and training a model to predict human ratings of 4 surface properties from touch. By contrast, we have significantly larger and more diverse data from indoor and outdoor scenes (rather than flat, largely indoor surfaces), use a rich vision-based tactile sensor (GelSight), and demonstrate our dataset on cross-modal prediction tasks.

Multimodal feature learning. In seminal work, de Sa [15] proposed to learn from correlating sight from sound. A variety of methods of been proposed for training deep networks to learn features from audio-visual correlations [49, 51, 52, 2, 36, 50, 47], from images and depth [62], and from vision and language [34, 45, 18, 56], and matching images and touch [39]. We adapt the contrastive model of Tian et al. [62] to visuo-tactile learning.

Multimodal image prediction. A variety of methods have been proposed for predicting images from another modality, such as by using text or labels [58, 31, 4, 57] or sound [40, 11, 42]. Li et al. [42] proposed a model for audio-driven stylization, i.e. learning to restyle images to better match an audio signal. We adapt this model to tactile-driven stylization, creating a model that is conditioned on tactile inputs instead of sound. We also take inspirations from work on future video prediction [17, 24, 54, 65, 21, 69, 66, 3, 63]. In particular, Tian et al. [61] trains an action-conditioned video prediction method to estimate future tactile signals, using a GelSight sensor controlled by a CNC machine. In contrast, we predict future tactile signals from natural objects, and show that visual information can improve the prediction quality.

Cross-modal image stylization. Many areas of multimodal perception have used cross-modal image stylization to evaluate whether models can capture associations between modalities (e.g. Text-to-image stylization, Audio-visual stylization). We adapt the cross-modal stylization method of Li et al. [42] to tactile-driven stylization, a task that requires learning visual-tactile associations – i.e., how an object feels from how it looks and vice versa. This direction is also related to work in computer graphics that synthesizes images that have specific material properties [41, 26, 27, 75]. In contrast to these works, we synthesize images with material properties that are captured implicitly from a touch signal.

3 The *Touch and Go* Dataset

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

3.1 Collecting a natural visuo-tactile dataset

To acquire our dataset, human data collectors (the authors) walked through a variety of environments, probing the objects with a tactile sensor. To obtain images that show clear, zoomed-in images of the objects being touched, two people collected data at once: one who presses the tactile sensor onto an object, and another who records an "approximately egocentric" video (see supplement for a visualization) of their hand. The two data collectors moved from object to object in the space as part of a single, continuously recording video, touching the objects around them. We show examples from our dataset in Fig. 2. The captured data varies heavily in material properties (e.g., soft/hard, smooth/rough), geometries, and semantics.
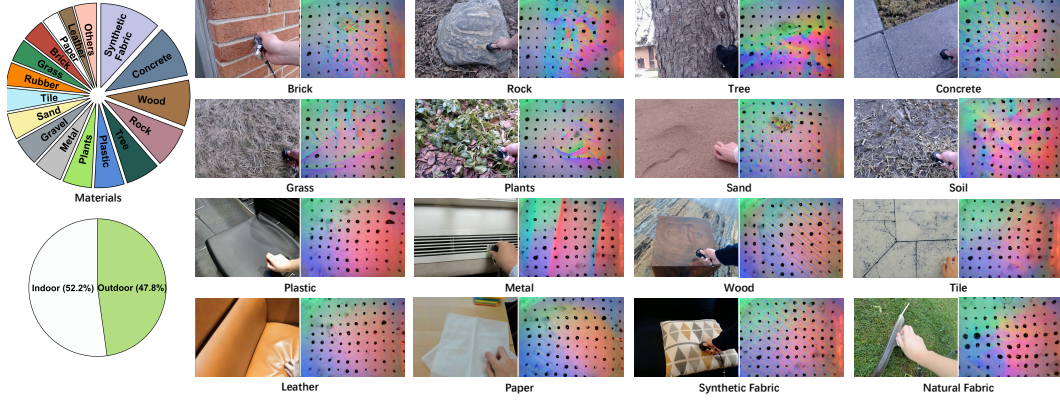
Figure 2: **The *Touch and Go* Dataset**. Human data collectors record paired visual and tactile information by probing objects in a variety of indoor and outdoor spaces. We show a selection of images, paired with the corresponding frame recorded by the GelSight tactile sensor. We show 16 representative categories (out of 20), and provide the distribution of material and scene types.[2]

Capturing procedure. To ensure that our dataset captures the natural variation of real-world vision and touch, we collect both rigid and deformable objects in indoor and outdoor scenes. These scenes include rooms in university buildings, such as classrooms and hallways, apartments, hiking trails, playgrounds, and streets. We show example footage from our model in Fig. 1, and in the supplement. The data collectors selected a variety of objects in each scene to press including chairs, walls, ground, sofa, table, etc. in indoor scenes, and grass, rock, tree, sand etc. in outdoor scenes, pressing each 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. Since the GelSight may provide information about force implicitly [71], and explicit force readings are not required for many visuo-tactile tasks, we do not use a separate force sensor.

Hardware. For the tactile sensor, we use GelSight [33], the variant designed for robotic manipulation [71]. 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. When the sensor is pressed against an object, the gel deforms, which results in changes to the reflected illumination. The color conveys the surface normal of the object being touched, similar to photometric stereo. These black dots are "markers" that are physically embedded within the GelSight's elastomer gel. Thus, 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 [28] that operates on GelSight images. On our test set of 2k hand-labeled frames, our classifier obtains 97% accuracy (chance is around 66%). To further ensure high accuracy, we hire workers to review the onset labels and correct errors.

---

[2]Following 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	Sensor
More Than a Feeling [8]	65	6.5k	Robot	✓	Tabletop	GelSight [70]
The Feeling of Success [9]	106	9.3k	Robot	✓	Tabletop	GelSight [70]
VisGel [43]	195	12k	Robot	✓	Tabletop	GelSight [70]
ObjectFolder 1.0 [22]	100	-	Synthetic	✗	-	DIGIT [37]
ObjectFolder 2.0 [23]	1000	-	Synthetic	✗	-	GelSight [70]
Burka et al. [7]	357	1.1k	Human	✓	Mostly Indoor	Multiple Sensors
Touch and Go (Ours)	3971	13.9k	Human	✓	Indoor/Outdoor	GelSight [70]

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 [51]. 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 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 [43] 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, 43] are confined to a single lab space containing the robot.

Qualitative comparison to other datasets We show qualitative examples of data from other datasets in Fig. 3 to help understand the differences between our dataset and those of previous work: Object Folder 2.0 [23], which contains virtual objects, and two robotic datasets: Feeling of Success [9], and VisGel [43]. We show examples from indoor scenes, since the other datasets do not contain outdoor scenes, and with rigid materials (since the virtual scenes do not contain deformable materials). Each row illustrates objects which are composed of similar materials, along with their corresponding GelSight images. As can be seen, the robot-centric datasets [43, 9] are confined to a fixed space. Their objects are also smaller than those in our dataset, since they must be capable of being grasped by the robot's gripper. Synthetic datasets [22, 23] do not contain complex microgeometry, and their rigid objects do not deform when pressed.

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
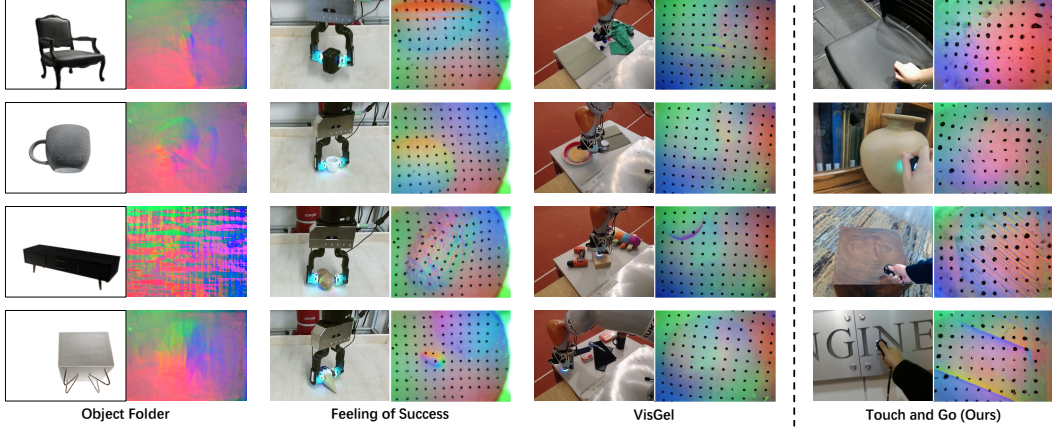
Figure 3: **Visuo-tactile data from other datasets**. We provide qualitative examples of visual and tactile data from other datasets (left), along with examples from similar material taken from our dataset (right).[3]

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. [62]. 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:

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

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:

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

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

4.2 Tactile-driven image stylization

Touch provides complementary information that may not be easily conveyed through other modalities that are commonly used to drive image stylization, such as language and sound. For example, touch can precisely define how smooth/rough a surface ought to be, and express the subtle shape of its microgeometry. A model that can successfully predict these properties from visuo-tactile data therefore ought to be able to translate between modalities. Inspired by the audio-driven image stylization of Li et al. [42], we propose the task of *tactile-driven image stylization*: making an image look as though it "feels like" a given touch signal.

Following Li et al. [42], we base our approach on contrastive unpaired translation (CUT) [53]. Given an input image $\mathbf{x}_I$ and a tactile example $\mathbf{x}_T$, our goal is to manipulate the image via a generator 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 consists of an image translation network that is conditioned on a tactile example (a GelSight image). The loss function encourages the model to preserve the image structure, which is enforced using an image-based contrastive loss [53], while adjusting the image style such that the resulting textures are more likely to co-occur with the tactile signal, which is measured using a visuo-tactile discriminator.

Table 2: Comparison of pretrained models on ImageNet and other tactile datasets on different downstream tasks.

Dataset	Method	Grasping Acc(%)	Material Acc(%)	Hard/Soft Acc (%)	Rough/Smooth Acc (%)
ImageNet [16]	Supervised	73.0	33.4	72.3	76.3
Object Folder 2.0 [23]	Visuo-tactile CMC	66.7	25.1	66.1	65.9
VisGel [43]	Visuo-tactile CMC	75.6	31.9	69.4	70.4
Ours - 25% Categories	Visuo-tactile CMC	62.3	23.9	67.3	65.3
Ours - 50% Categories	Visuo-tactile CMC	66.5	29.1	71.2	66.8
Ours - 75% Categories	Visuo-tactile CMC	70.8	32.5	73.3	74.7
Ours - 100% Categories	Visuo-tactile CMC	**78.6**	**35.5**	**77.3**	**79.4**

Making sight consistent with touch. We train the model with a discriminator D that distinguishes between real (and fake) visuo-tactile pairs. During training time, we shuffle the dataset to generate the set S_n containing mismatched image-tactile pairs $\{\mathbf{x}_I, \mathbf{x}'_T\} \in S_n$. Likewise, we define the original dataset as S_m which contains matched image-tactile pairs $\{\mathbf{x}_I, \mathbf{x}_T\} \in S_m$. In formal terms, the visual-tactile adversarial loss can be written as:

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

Loss. We combine our visuo-tactile discriminator with the structure-preserving loss used in Li et al. [43], which was originally proposed by CUT [53]. This loss, which we call $\mathcal{L}_{\text{CUT}}$, works by training a contrastive learning model that puts patches in the input and predicted images into correspondence, such that patches at the same positions are close in an embedding space. Please see the supplement for more details. The overall loss is:

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

4.3 Multimodal future touch prediction

Inspired by the challenges of action-conditional future touch prediction [61], we use our dataset to ask whether *visual* data can improve our estimates of future tactile signals: i.e., what will this object feel like in a moment? Visual information can help touch prediction in a number of ways, such as by conveying material properties (e.g., deformation) and and geometry. It can also provide information about which action is being performed. One may thus also consider it as an *implicit* form of action conditioning [20], since the visual information provides information analogous to actions.

We adapt the video prediction architecture of Geng et al. [24] to this task. This model is based on residual networks [28, 30] with 3D space-time convolutions (please see the supplement for architecture details). We predict multiple frames by autoregressively feeding our output images back to the original model. Given a series of paired visual and tactile images from times 1 to t, $\{(\mathbf{x}_I^1, \mathbf{x}_T^1), ..., (\mathbf{x}_I^t, \mathbf{x}_T^t)\}$, our goal is to predict the subsequent tactile image, $\hat{\mathbf{x}}_T^{(t+1)}$. We train the model with L1 and perceptual loss, following [24].

5 Experiments

5.1 Self-supervised feature learning

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

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

Robotic grasping task. For the robot grasping task, we use the experimental setup and dataset of Calandra et al. [9]. Thus, the task requires generalizing to data recorded in a very different environment, and with different GelSight sensors. The goal of this task is to predict whether a robotic arm will successfully grasp an object, based on inputs from two GelSight images recorded 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 understanding tasks.   We evaluate whether the learned features convey material properties. Given tactile features, we recognize: 1) material categories, 2) hard vs. soft surfaces, and 3) smooth vs. rough surfaces. Following [51], we re-categorize material categories to generate soft/hard labels and hire online workers to label the smooth/rough according to the visual image. Since the smoothness and roughness may vary within a material category, we hire online workers to label the smooth vs. rough according to the visual 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 frame) to avoid having the same (or nearly same) object appear in both training and test.

Implementation details.   We train our model for 240 epochs, using the optimization parameters from CMC [62], 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 [62] 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 [72, 9, 8], and visual CMC [62] 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. We also show that increasing the material categories leading to much better downstream performance.

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. [42], 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 [29]. The architecture of discriminator follows [42], 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 [42], 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 [42], we use our trained

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

Method	CMC	Material
Baseline	0.165	0.107
CycleGAN [76]	0.178	0.129
Ours	**0.197**	**0.142**

CMC model, by taking the dot product between visual and tactile embeddings. Second, we compare 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 [76]. This model performs the stylization using CycleGAN, rather than CUT (see supplement for details).

Figure 4: Qualitative results of 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.*

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 [76]. In terms of material classification consistency, our model consistently outperforms CycleGAN-based method [76].

Qualitative results. In Fig. 4, 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.

5.3 Multimodal video prediction

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

9

Table 4: Quantitative results for video prediction.

Time horizon	Method	Modality	L1 ↓	SSIM ↑	LPIPS ↓
Skip 3 frames	SVG [17]	Touch only	0.782	0.572	0.391
	Geng et al. [24]	Touch only	0.628	0.708	0.103
	SVG [17]	Touch + vision	0.757	0.602	0.368
	Geng et al. [24]	Touch + vision	**0.617**	**0.719**	**0.091**
Skip 5 frames	SVG [17]	Touch only	0.807	0.513	0.412
	Geng et al. [24]	Touch only	0.691	0.698	0.279
	SVG [17]	Touch + vision	0.762	0.546	0.397
	Geng et al. [24]	Touch + vision	**0.663**	**0.713**	**0.265**
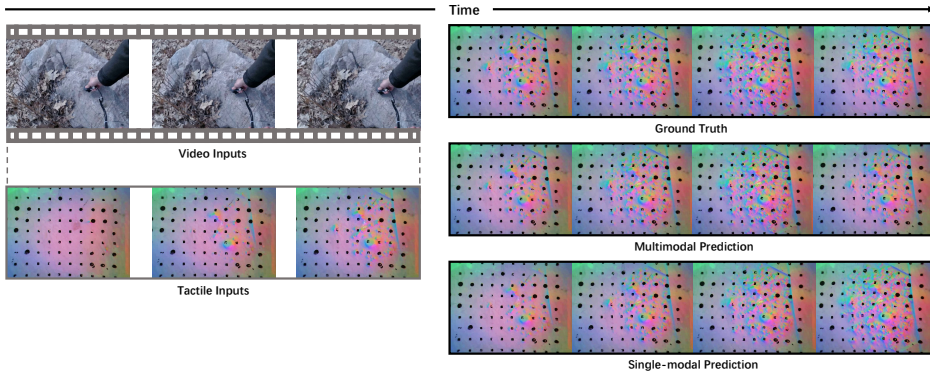
Figure 5: **Future touch prediction.** We show the results of tactile-only and visuo-tactile models.

Experimental setting. We evaluate the effectiveness of multimodal inputs using three context frames to predict the next frame under two different time horizons: skipping 3 and 5 frames between contexts. Following [24], we adopt three evaluation metrics: MAE, SSIM [67] and LPIPS [74]. We provide training hyperparameters in the supplement.

Multimodal vs. single-modal prediction. We show the quantitative results in Table 4. Here, we adopt two video prediction baselines [17] and [24]. We can see that, by incorporating our dataset's visual signal, the models gain a constant performance increase under different evaluation metrics for both model, under both experimental settings. The gap becomes larger for longer time horizon, suggesting that visual information may be more helpful in this case.

6 Discussion

We proposed *Touch and Go*, a human-collected visuo-tactile dataset. Our dataset comes from real-world objects, and is significantly more diverse than prior datasets. We demonstrate its effectiveness on a variety of applications that involve robotic manipulation, material understanding, and image synthesis. In the tradition of previous work [51], we see our work as a step toward *human-collected* multimodal data collection, in which humans equipped with multiple sensors collect diverse dataset by recording themselves physically interacting with the world. We hope this data will enable researchers to study diverse visuo-tactile learning applications, beyond the "robotics-centric" domains that are often the focus of previous efforts.

Limitations. Collecting diverse tactile data is an ongoing challenge, since it requires physically being present in the locations where data is collected. While adding human collectors improves diversity in many ways, our dataset was mainly collected in one geographic location (near University of Michigan's campus). Consequently, the data we recorded may not generalize to all spaces. The use of humans in the data collection process also potentially introduces bias, which differs from "robotic" or "virtual data" bias. For example, humans may choose unrepresentative parts of the objects to probe, and do not perform actions with consistent force. The humans who recorded the dataset may also not be representative of the general population, which may introduce bias (e.g., in skin tone).

References

[1] Edward H Adelson. On seeing stuff: the perception of materials by humans and machines. In *Human vision and electronic imaging VI*, volume 4299, pages 1–12. SPIE, 2001.

[2] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 609–617, 2017.

[3] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H Campbell, and Sergey Levine. Stochastic variational video prediction. *arXiv preprint arXiv:1710.11252*, 2017.

[4] David Bau, Alex Andonian, Audrey Cui, YeonHwan Park, Ali Jahanian, Aude Oliva, and Antonio Torralba. Paint by word. In *arXiv:2103.10951*, 2021.

[5] Alex Burka, Siyao Hu, Stuart Helgeson, Shweta Krishnan, Yang Gao, Lisa Anne Hendricks, Trevor Darrell, and Katherine J Kuchenbecker. Proton: A visuo-haptic data acquisition system for robotic learning of surface properties. In *2016 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, pages 58–65. IEEE, 2016.

[6] Alex Burka, Abhinav Rajvanshi, Sarah Allen, and Katherine J Kuchenbecker. Proton 2: Increasing the sensitivity and portability of a visuo-haptic surface interaction recorder. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 439–445. IEEE, 2017.

[7] Alexander L Burka. Instrumentation, data, and algorithms for visually understanding haptic surface properties. *Dissertations available from ProQuest*, 2018.

[8] Roberto Calandra, Andrew Owens, Dinesh Jayaraman, Justin Lin, Wenzhen Yuan, Jitendra Malik, Edward H. Adelson, and Sergey Levine. More than a feeling: Learning to grasp and regrasp using vision and touch. *IEEE Robotics and Automation Letters*, 3:3300–3307, 2018.

[9] Roberto Calandra, Andrew Owens, Manu Upadhyaya, Wenzhen Yuan, Justin Lin, Edward H Adelson, and Sergey Levine. The feeling of success: Does touch sensing help predict grasp outcomes? *Conference on Robot Learning (CoRL)*, 2017.

[10] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[11] Moitreya Chatterjee and Anoop Cherian. Sound2sight: Generating visual dynamics from sound and context. In *European Conference on Computer Vision*, pages 701–719. Springer, 2020.

[12] Vivian Chu, Ian McMahon, Lorenzo Riano, Craig G McDonald, Qin He, Jorge Martinez Perez-Tejada, Michael Arrigo, Trevor Darrell, and Katherine J Kuchenbecker. Robotic learning of haptic adjectives through physical interaction. *Robotics and Autonomous Systems*, 63:279–292, 2015.

[13] Shaowei Cui, Rui Wang, Junhang Wei, Jingyi Hu, and Shuo Wang. Self-attention based visual-tactile fusion learning for predicting grasp outcomes. *IEEE Robotics and Automation Letters*, 5(4):5827–5834, 2020.

[14] Mark R. Cutkosky, Robert D. Howe, and William R. Provancher. *Force and Tactile Sensors*, pages 455–476. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[15] Virginia R de Sa. Learning classification with unlabeled data. *Advances in neural information processing systems*, 1994.

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[17] Emily L. Denton and Rob Fergus. Stochastic video generation with a learned prior. In *ICML*, 2018.

[18] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11162–11173, 2021.

[19] Nima Fazeli, Miquel Oller, Jiajun Wu, Zheng Wu, Joshua B Tenenbaum, and Alberto Rodriguez. See, feel, act: Hierarchical learning for complex manipulation skills with multisensory fusion. *Science Robotics*, 4(26):eaav3123, 2019.

[20] Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. *Advances in neural information processing systems*, 29, 2016.

[21] Hang Gao, Huazhe Xu, Qi-Zhi Cai, Ruth Wang, Fisher Yu, and Trevor Darrell. Disentangling propagation and generation for video prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9006–9015, 2019.

[22] Ruohan Gao, Yen-Yu Chang, Shivani Mall, Li Fei-Fei, and Jiajun Wu. Objectfolder: A dataset of objects with implicit visual, auditory, and tactile representations. In *CoRL*, 2021.

[23] Ruohan Gao, Zilin Si, Yen-Yu Chang, Samuel Clarke, Jeannette Bohg, Li Fei-Fei, Wenzhen Yuan, and Jiajun Wu. Objectfolder 2.0: A multisensory object dataset for sim2real transfer. In *CVPR*, 2022.

[24] Daniel Geng, Max Hamilton, and Andrew Owens. Comparing correspondences: Video prediction with correspondence-wise losses. *Computer Vision and Pattern Recognition (CVPR)*, 2022.

[25] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[26] Paul Guerrero, Milos Hasan, Kalyan Sunkavalli, Radomir Mech, Tamy Boubekeur, and Niloy Mitra. Matformer: A generative model for procedural materials. *ACM Trans. Graph.*, 41(4), 2022.

[27] Yu Guo, Cameron Smith, Miloš Hašan, Kalyan Sunkavalli, and Shuang Zhao. Materialgan: Reflectance capture using a generative svbrdf model. *ACM Trans. Graph.*, 39(6):254:1–254:13, 2020.

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[29] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, 2017.

[30] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution, 2016.

[31] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1219–1228, 2018.

[32] Micah Johnson, Forrester Cole, Alvin Raj, and Edward Adelson. Microgeometry capture using an elastomeric sensor. *ACM Trans. Graph.*, 30:46, 07 2011.

[33] Micah K Johnson and Edward H Adelson. Retrographic sensing for the measurement of surface texture and shape. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1070–1077. IEEE, 2009.

[34] Armand Joulin, Laurens van der Maaten, Allan Jabri, and Nicolas Vasilache. Learning visual features from large weakly supervised data. In *European Conference on Computer Vision*, pages 67–84. Springer, 2016.

[35] Zhanat Kappassov, Juan-Antonio Corrales, and Véronique Perdereau. Tactile sensing in dexterous robot hands — review. *Robotics and Autonomous Systems*, 74:195–220, 2015.

[36] Bruno Korbar, Du Tran, and Lorenzo Torresani. Co-training of audio and video representations from self-supervised temporal synchronization. *CoRR*, 2018.

[37] Mike Lambeta, Po wei Chou, Stephen Tian, Brian Yang, Benjamin Maloon, Victoria Rose Most, Dave Stroud, Raymond Santos, Ahmad Byagowi, Gregg Kammerer, Dinesh Jayaraman, and Roberto Calandra. Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robotics and Automation Letters*, 5:3838–3845, 2020.

[38] Susan J Lederman and Roberta L Klatzky. Hand movements: A window into haptic object recognition. *Cognitive Psychology*, 19(3):342–368, 1987.

[39] Michelle A Lee, Yuke Zhu, Krishnan Srinivasan, Parth Shah, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Jeannette Bohg. Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. *arXiv preprint arXiv:1810.10191*, 2018.

[40] Seung Hyun Lee, Wonseok Roh, Wonmin Byeon, Sang Ho Yoon, Chan Young Kim, Jinkyu Kim, and Sangpil Kim. Sound-guided semantic image manipulation. *arXiv preprint arXiv:2112.00007*, 2021.

[41] Hendrik PA Lensch, Michael Goesele, Yung-Yu Chuang, Tim Hawkins, Steve Marschner, Wojciech Matusik, and Gero Mueller. Realistic materials in computer graphics. In *ACM SIGGRAPH 2005 Courses*, pages 1–es. 2005.

[42] Tingle Li, Yichen Liu, Andrew Owens, and Hang Zhao. Learning visual styles from audio-visual associations. *European Conference on Computer Vision (ECCV)*, 2022.

[43] Yunzhu Li, Jun-Yan Zhu, Russ Tedrake, and Antonio Torralba. Connecting touch and vision via cross-modal prediction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[44] Justin Lin, Roberto Calandra, and Sergey Levine. Learning to identify object instances by touch: Tactile recognition via multimodal matching. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 3644–3650. IEEE, 2019.

[45] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 181–196, 2018.

[46] Andrew Melnik, Luca Lach, Matthias Plappert, Timo Korthals, Robert Haschke, and Helge Ritter. Tactile sensing and deep reinforcement learning for in-hand manipulation tasks. In *IROS Workshop on Autonomous Object Manipulation*, 2019.

[47] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12475–12486, 2021.

[48] Adithyavairavan Murali, Yin Li, Dhiraj Gandhi, and Abhinav Gupta. Learning to grasp without seeing. In *International Symposium on Experimental Robotics*, pages 375–386. Springer, 2018.

[49] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *ICML*, 2011.

[50] Andrew Owens and Alexei A. Efros. Audio-visual scene analysis with self-supervised multisensory features. *Eur. Conf. Comput. Vis.*, 2018.

[51] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H. Adelson, and William T. Freeman. Visually indicated sounds. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2405–2413, 2016.

[52] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *Eur. Conf. Comput. Vis.*, 2016.

[53] Taesung Park, Alexei A. Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for conditional image synthesis. In *ECCV*, 2020.

[54] Viorica Patraucean, Ankur Handa, and Roberto Cipolla. Spatio-temporal video autoencoder with differentiable memory. *arXiv preprint arXiv:1511.06309*, 2015.

[55] Senthil Purushwalkam, Abhinav Gupta, Danny M Kaufman, and Bryan Russell. Bounce and learn: Modeling scene dynamics with real-world bounces. *arXiv preprint arXiv:1904.06827*, 2019.

[56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[57] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021.

[58] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, pages 1060–1069, 2016.

[59] Philipp Ruppel, Yannick Jonetzko, Michael Görner, Norman Hendrich, and Jianwei Zhang. Simulation of the syntouch biotac sensor. In *International Conference on Intelligent Autonomous Systems*, pages 374–387. Springer, 2018.

[60] Subramanian Sundaram, Petr Kellnhofer, Yunzhu Li, Jun-Yan Zhu, Antonio Torralba, and Wojciech Matusik. Learning the signatures of the human grasp using a scalable tactile glove. *Nature*, 569(7758):698–702, 2019.

[61] Stephen Tian, Frederik Ebert, Dinesh Jayaraman, Mayur Mudigonda, Chelsea Finn, Roberto Calandra, and Sergey Levine. Manipulation by feel: Touch-based control with deep predictive models. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 818–824. IEEE, 2019.

[62] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pages 776–794. Springer, 2020.

[63] Ruben Villegas, Arkanath Pathak, Harini Kannan, Dumitru Erhan, Quoc V. Le, and Honglak Lee. High fidelity video prediction with large stochastic recurrent neural networks, 2019.

[64] Shaoxiong Wang, Mike Lambeta, Po-Wei Chou, and Roberto Calandra. Tacto: A fast, flexible, and open-source simulator for high-resolution vision-based tactile sensors. *IEEE Robotics and Automation Letters*, 7(2):3930–3937, 2022.

[65] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. Few-shot video-to-video synthesis. *arXiv preprint arXiv:1910.12713*, 2019.

[66] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans, 2017.

[67] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[68] Youngsun Wi, Pete Florence, Andy Zeng, and Nima Fazeli. Virdo: Visio-tactile implicit representations of deformable objects. *arXiv preprint arXiv:2202.00868*, 2022.

[69] Yue Wu, Rongrong Gao, Jaesik Park, and Qifeng Chen. Future video synthesis with object motion prediction. In *CVPR*, 2020.

[70] Wenzhen Yuan, Siyuan Dong, and Edward Adelson. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17:2762, 11 2017.

[71] Wenzhen Yuan, Siyuan Dong, and Edward H Adelson. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17(12):2762, 2017.

[72] Wenzhen Yuan, Chenzhuo Zhu, Andrew Owens, Mandayam A Srinivasan, and Edward H Adelson. Shape-independent hardness estimation using deep learning and a gelsight tactile sensor. In *International Conference on Robotics and Automation (ICRA)*, 2017.

[73] Qiang Zhang, Yunzhu Li, Yiyue Luo, Wan Shou, Michael Foshey, Junchi Yan, Joshua B Tenenbaum, Wojciech Matusik, and Antonio Torralba. Dynamic modeling of hand-object interactions via tactile sensing. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2874–2881. IEEE, 2021.

[74] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018.

[75] Xilong Zhou and Nima Khademi Kalantari. Adversarial single-image svbrdf estimation with hybrid training. In *Computer Graphics Forum*, volume 40, pages 315–325. Wiley Online Library, 2021.

[76] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] No potential negative social impacts.

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...

   (a) Did you state the full set of assumptions of all theoretical results? [N/A] No theoretical results

   (b) Did you include complete proofs of all theoretical results? [N/A] No theoretical results

3. If you ran experiments (e.g. for benchmarks)...

   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]

   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they 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.