TOUCH and GO: A Real-World Multisensory Dataset

Chenyang Ma[*]
University of Michigan
dannymcy@umich.edu

Fengyu Yang[*]
University of Michigan
fredyang@umich.edu

Andrew Owens
University of Michigan
ahowens@umich.edu

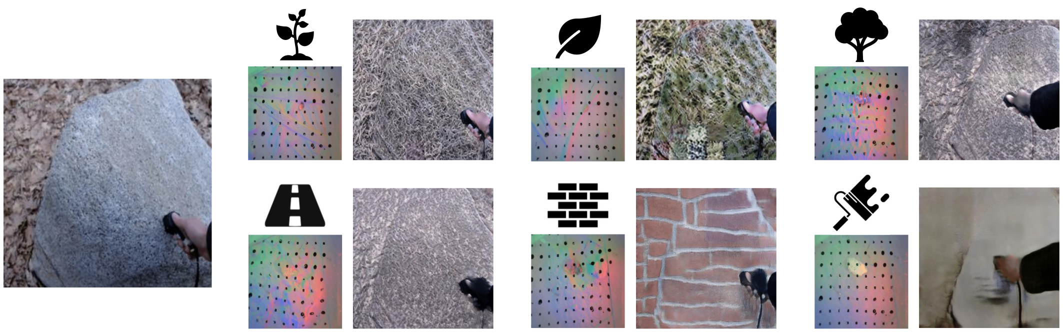

Figure 1: *Tactile-Guided Image Stylization* (TGIS). We present a method for manipulating the appearance of an object to match its material property indicated by the tactile signal.

Abstract

Unlike how humans perceive the world from associations between senses and through a series of inanimate objects, contemporary research on robot perception problem mainly rely on vision units or visual inputs to teach the robots interact with the world. We identify that this is due to the lack of real-world multisenory rich object dataset. To tackle this challenge, we present **TOUCH and GO**, a multisensory dataset containing real-world synchronized high-quality video and tactile data containing 12600 object instances over 37800 touches and 30 hours of video captured from egocentric viewpoint, greatly exceeding the size of existing real-world multisensory datasets. All objects in our dataset are originated from real environments with fine-grained textures retained. We propose and apply our dataset on two novel tasks, tactile-guided image stylization and multi-modal video prediction on tactile images.

1 Introduction

Humans perceive the world not using a single modality. Instead, we have access to many sensory streams and learn from associations between senses. When a child eats an apple, for instance, she'll not only taste it—she'll also hear it crunch, see its shiny skin, and feel its smooth surface [48]. In addition, humans perceive the world not as a single giant entity but often through a series of inanimate objects, which exist as bounded wholes and move on connected paths. We interact with these objects

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

Submitted to the 36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks. Do not distribute.

through an array of different sensory systems–vision, touch, audition, smell, taste, and proprioception. These multisensory inputs shape our daily experiences.

It is so nature for humans to learn knowledge through interactions with different objects with multiple senses. Cognitive science studies [49, 48] show that both object representation and multisensory perception play a crucial role in early human cognitive development. However, for robots, it may not be the case. Contemporary research on robot perception problem mainly rely on vision units or visual inputs to teach the robots perceive and interact with the world. This focus on learning from vision alone makes the perception problem harder because some of the most important spectrum of physical object properties and sensory modes — such as touching — are lost.

We identify that this is due to the lack of real-world multisenory rich object dataset. Several works have been done regarding simulated multisensory dataset [16, 17, 14]. However, we argue there are two fundamental differences between the quality and utility of simulated and real-world dataset. First, simulated data fail to perfectly represent reality. Models trained purely on synthetic data do not generalize to the real world due to the discrepancy between simulated and real environments, in terms of both visual and physical properties. In fact, the more we increase the fidelity of our simulations, the more effort we have to expend in order to build them, both in terms of implementing complex physical phenomena and in terms of creating the content (e.g., objects, backgrounds) to populate these simulations. This difficulty is compounded by the fact that powerful optimization methods based on deep learning are exceptionally proficient at exploiting simulator flaws: the more powerful the machine learning algorithm, the more likely it is to discover how to "cheat" the simulator to succeed in ways that are infeasible in the real world [4]. Second, "reality gap" exits by transferring simulated experience into the real world. While simulated data continue to improve in fidelity, the peculiar and pathological regularities of synthetic data, and the wide, unpredictable diversity of real-world objects, makes bridging the reality gap particularly difficult when the robot use its sensors to perceive the world, as is the case for example in many manipulation tasks [4, 29].

Therefore, our goal is to establish a real-world multisensory dataset contraining rich objects that are 1) easily accessible to the community as a standard benchmark, 2) high-quality in terms of visual textures, and 3) augmented with real data from the perspective of human beings. To this end, we introduce TOUCH and GO — an egocentric multisensory dataset of synchronized video and tactile sensing. We take inspiration from the way infants explore the physical properties of a scene by poking and prodding at the objects in front of them [3, 46], a process that may help them learn an intuitive theory of physics. The egocentric viewpoint enables our dataset to contain enough details to observe the fine-grained texture of objects, and mimics the perception of a real human.

More specifically, we collect over 30 hours of real-world synchronized high-quality video and tactile data containing 12600 object instances over 37800 touches. Our dataset contains rich objects categories from both indoor and outdoor scenes (none of the existing real-world multi-modal dataset contains data from outdoor scenes). TOUCH and GO enables many applications. We present a method for manipulating the appearance of an object to match its material property indicated by the tactile signal, a problem we term *Tactile-Guided Image Stylization* (TGIS), as shown in Figure 1. We also propose a novel multi-modal video prediction problem on tactile image deformation. For the first task, We design a deep neural network based on CUT [44], which fuses data from video and tactile streams. For both tasks, experimental results suggest better results are achieved by leveraging our TOUCH and GO dataset.

Our main contributions can be concluded as the followings: 1) We introduce TOUCH and GO, a real-world dataset that makes multisensory learning with vision and touch easily accessible to the research community. 2) All objects in our dataset are originated from real environments and will be made publicly available as a standard testbed for robotic multisensory learning. 3) We propose and apply our dataset on two novel tasks including tactile-guided image stylization and multi-modal video prediction on tactile images.

2 Related Work

Multisensory Datasets There is a mixture of real and simulated data across different single-modal datasets. ImageNet [12], MS COCO [35], ObjectNet [1], and OpenImages [30] focus on the collection of large-scale real 2D images. ModelNet [56] and ShapeNet [9] contain synthetic 3D CAD models, emphasizing on geometry of 3D objects but pay less attention to fine-grained visual textures. BigBIRD [47], YCB [8], and ABO [10] model real-world 3D objects with limited object instances. The majority of multi-modal datasets incorporate simulated data. Pix3D [51], IKEA Objects [34], and Object3D [57] match synthetic 3D CAD models to objects in real images. OBJECT-FOLDER 1.0 [16] contains multisensory simulated data as implict neural representations. Built upon it, OBJECT-FOLDER 2.0 [17] is ten times larger than the previous version with encoding of more realistic data. A few real-world multi-modal datasets exist. VisGel [33] comprises real-world data of videos and touches collected by robotics arm, thus has very restricted scenes and bias introduced by the arm. Greatest Hits [43] contains high-quality egocentric videos of humans probing environments with a drumstick, but its goal is not on scale expansion and generalization. Our TOUCH and GO dataset contains high-quality synchronized RGB video and tactile data, with over 30 hours of videos, 37800 touches, and 12600 object instances, which greatly exceeds the size of existing real-world egocentric datasets.

Touch and Vision Researches are conducted on the types of haptic, force, and tactile sensors to give robots tactile sensing ability [11, 25, 32, 31]. GelSight [23, 22, 59, 7] is widely adapted as a high-resolution tactile sensor for computer vision and robotics applications, which includes improving grasp stability with rotation measurement [28], the study of the physical and material properties of fabrics [60], predicting the grasping success through both vision and tactile sensing [7], and cloth texture recognition [38]. Here we introduce the novel application of tactile-guided image stylization.

Image-to-Image Stylization Image stylization (translation) translates an input image from one domain to a photo realistic output in the target domain [20, 36, 61]. The key to the success of this task is due to the emergence of generative adversarial networks (GAN) [18, 41], which have been vigorously researched in the last several years with many applications including generating photos from sketches [20, 45], changing time of a day [20, 63], and translating semantic meanings into scenes [20, 55]. While most of the image stylization tasks have paired image-to-image translation, in certain cases, the corresponding examples from domains are unavailable, resulting in unpaired image-to-image stylization. Cycle consistency [26, 58, 62], as one of the approaches, enforce the correspondence between the input and output image domain by adopting the underlying bijective assumption, which may be too restrictive in cases when images from one domain contain additional information compared to the other domain. CUT [44] adapts contrastive learning to make each patch in the output reflect the content of the corresponding patch in the input by maximizing mutual information between the two. We propose a new model based on CUT, which receives multi-modal data as inputs and learns to build the tactile-visual style associations without any human supervison, for our proposed novel task of tactile-guided image stylization.

Video Prediction Approaches for video prediction are diverse, evolving from the modeling of long-range dependencies recurrent networks [24, 40, 42, 50, 52, 5] to photorealistic video prediction using large convolutional neural networks [37, 54, 39]. Time-agnostic prediction [21], which enables model to predict any future frames in a video, is also proposed. In addition, methods based variational autoencoders (VAEs) [27, 2, 15, 13, 53] are introduced to tackle the challenges of uncertainty in video prediction. Our approach uses VAE-based video prediction model [13, 53] to combine multi-modal data as inputs and predict the next frame tactile images.

3 TOUCH and GO Dataset

We collect a real-world vision-tactile dataset that contains egocentric videos of human (the authors) pressing environments using a tactile sensor, Gelsight, and the tactile information from the Gelsight that is simultaneously recorded with the RGB video. The touch of the environment contains useful information about an object associated with the visual information, including hardness, shape,

Figure 2: TOUCH and GO Dataset. What do these objects feel like when they are touched? Here, we show some images from a selection of videos from our dataset for a subset of the object instances.

Table 1: Comparison of touch datasets.

	Hours	Touches	Object Inst.	Real-World	Indoor	Outdoor
More Than a Feeling [6]	-	6450	65	✓	✓	✗
VisGel [33]	20-30	12000	195	✓	✓	✗
The Feeling of Success [7]	-	9269	106	✓	✓	✗
Object Folder [16]	-	-	100	✗	-	-
Object Folder 2.0 [17]	-	-	1000	✗	-	-
TOUCH and GO (Ours)	>30	>37800*	>12600*	✓	✓	✓

material etc, which can be useful in various downstream tasks. Unlike traditional scene-centric datasets focusing on the full scene, our dataset is taken from an egocentric viewpoint which contains enough details to observe the fine-grained texture of an object.

3.1 Dataset Description

We collect over 30 hours of videos consisting of over 37800 touches from over 12600 objects under both indoor (58%) and outdoor (42%) scenes. In total, there are 1.89M frames containing touches of an object and each touch is composed of 50 frames on average. Our dataset contains daily seen objects, both hard and deformable, from indoor and outdoor scenes including rock, grass, road, brick, carpet, chair, table and so on. All the touch frames are annotated with the name of the object.

3.2 Comparison with other datasets

We compare TOUCH and GO with existing multisensory datasets in Table 1. Compared to the largest real-world dataset collected by robot, VisGel [33], our dataset comprises of longer hours of video, more touches, and most importantly much more diverse of object instances, where the total object instances is 65 times larger. It is worth noting that VisGel [33] and other robot collected datasets only contain indoor scene and the background is mostly fixed to the robotic operating station, which is far from the place where the object actually exists in the real world. Our dataset is collected by human from exact the real world where each object is recorded under the natural environment. With respect to the largest object-centric multisensory dataset Object Folder 2.0 [17], our dataset contains object instances 10 times larger and all tactile inputs are completely recorded by touching the actual object, which provides more realistic tactile data compared to the synthetic or simulated images.

3.3 Data Collection Setup

As shown in Figure 2, we utilize a webcam to record the RGB video and a GelSight sensor to capture the tactile signals, which are both connected to our PC. We record the timestamp of each frame to synchronize visual and tactile images. GelSight sensor [23, 22, 59, 7] is an optical tactile sensor that enables high spatial resolution measurement of the texture and geometry of a contact surface. The sensor consists of a 1.5cm × 1.5cm surface of a soft elastomer painted with a reflective membrane, which deforms to the shape of the object upon contact. There exists an ordinary camera beneath the elastomer so that we can view the deformed gel. The gels are illuminated by colored LEDs from different directions, producing a three-channel surface normal image. Thus, we can observe the texture of a surface undergoing the deformation process via consecutive 2D images. We can then treat the tactile images as normal 2D images, and pass them to visual backbone network to extract tactile information.

3.4 Detecting Touch Onset

According to our dataset, we have approximately 1/3 of the video frames that the GelSight sensor is not touching the object. This is because our dataset is collected by human moving around and touching objects seen during the movement. Thus the GelSight sensor has no deformation during the interval when human is moving from one object to another. However, at the mean time, the RGB camera still records the scene during the video. Under this circumstance, the scenes captured by RGB camera will be incorrectly linked to the tactile signal of no deformation, which will negatively impact the downstream tasks. Thus, to alleviate this issue, we train a binary classifier to classify whether the frame is at touch onset. We hand label 10,000 frames from the dataset and finetune ResNet-18 [19] initialized by weights pretrained on the ImageNet [12] on our dataset as our classifier. We report a 97% accuracy on our test set that is 20% of our labeled frames.

4 Applications

4.1 Tactile-Guided Image Stylization

The sense of touch conveys useful information about an object, including hardness, shape, material etc., which creates an inherent association with the visual input in the video. This connection between visual and tactile signals is embedded in our dataset and the neural network is able to build the tactile-visual style associations without human supervision. Moreover, tactile signals may provide us subtle distinction about objects that visual input can not capture. As shown in Figure 3, even when objects of different materials share similar visual appearance, the tactile signals are able to reveal their subtle difference. Given the unique properties of tactile signals and its association with visual input, we propose *Tactile-Guided Image Stylization* (TGIS) application on our dataset, which, to the best of our knowledge, is firstly considered in the current literature.

4.1.1 Proposed Method

Given an source domain $\mathcal{X} \in \mathbf{R}^{\mathbf{H}\times\mathbf{W}\times\mathbf{C}}$, our goal in TGIS is to learn the translation from $\mathcal{X}$ to look visually similar to an image from the target domain Y that is corresponded to the tactile domain $\mathcal{T}$. During the training time, we randomly sample two visual images from $\mathcal{X}$, $\mathcal{Y}$ and a tactile image from $\mathcal{T}$ corresponding to the target image $\mathcal{Y}$. It is worth noting that our training requires no human annotation and it can be done under self-supervision.

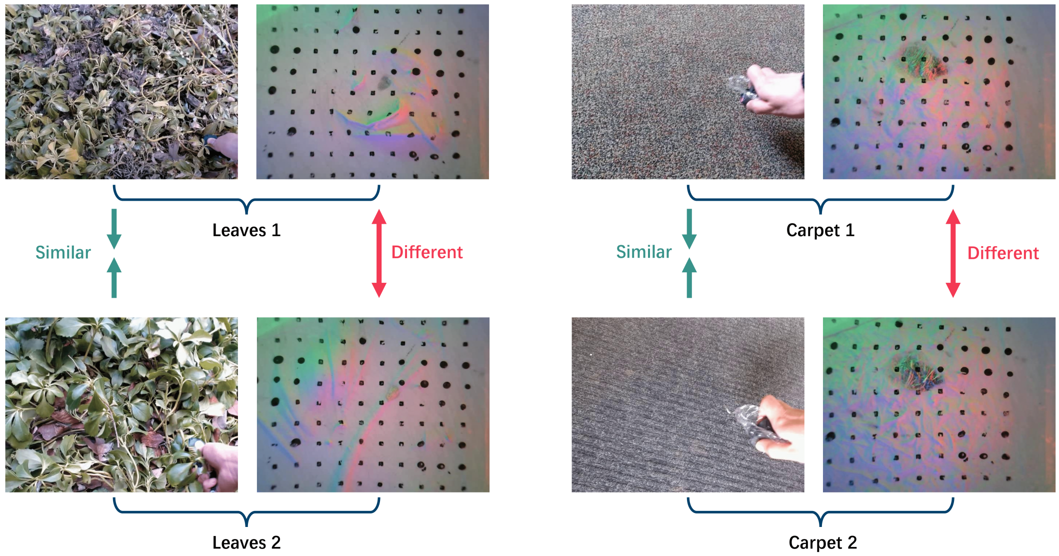

Figure 3: Subtle distinction from gelsight. Although it is hard to distinguish some object instances from visual appearance, tactile signals may convery enough information.

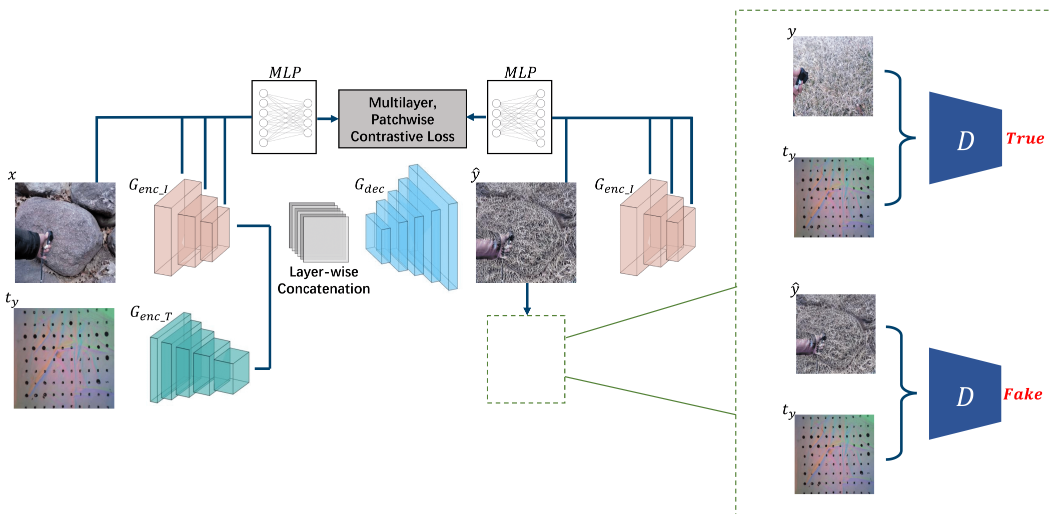

Figure 4: Pipeline of tactile-guided image stylization.

As shown in the Figure 4 about our pipeline, our model consists of a multi-modal generator, a tactile-visual texture discriminator, and a patch-wise structure discriminator. We can further break up our multi-modal generator into three components, an image encoder $G_{\mathrm{enc\_I}}$, a tactile encoder $G_{\mathrm{enc\_T}}$, and a decoder G_{dec}. Given our dataset that contains unpaired instances $X = \{\mathbf{x} \in \mathcal{X}\}, Y = \{\mathbf{y} \in \mathcal{Y}\}$ and tactile input $T_y = \{\mathbf{t_y} \in \mathcal{T}\}$, the output image $\hat{\mathbf{y}}$ can be expressed as $\hat{\mathbf{y}} = G(\mathbf{x}, \mathbf{t_y}) = G_{\mathrm{dec}}(\mathrm{concat}(\mathrm{G}_{\mathrm{enc\_I}}(\mathbf{x}), \mathrm{G}_{\mathrm{enc\_T}}(\mathbf{t_y})))$.

Tactile-Visual Adversarial Loss To leverage the association between visual input and tactile input, we propose a tactile-visual adversarial loss between $\hat{\mathbf{y}}$ and $\mathbf{t_y}$. In formal terms:

$$\mathcal{L}_{\mathrm{GAN}} D(\mathrm{G}_{X\to Y}, \mathrm{D}_Y) = \mathbb{E}_{\mathbf{y}\sim Y} \log D(\mathbf{y}, \mathbf{t_y}) + \mathbb{E}_{\mathbf{x}\sim X} \log(1 - D(G(\mathbf{x}, \mathbf{t_y}), \mathbf{t_y})) \tag{1}$$

where D is the discriminator. For the discriminator D, we adopt the early fusion where we first directly concatenate the generated image $\hat{\mathbf{y}}$ with the tactile input $\mathbf{t_y}$ and then feed into the discriminator D.

Structure Preservation via Contrastive Learning Our goal in this tactile-guided image stylization is to restyle the source image with the textures that are associated with the target tactile input while preserving the source structure. However, structure and texture of an image are often entangled with each other. With only tactile-visual adversarial loss, it becomes a trivial solution to completely transfer the source image to the target domain without preserving the original structure. Thus, we introduce an additional constraint called noise contrastive estimation (NCE) [44] to preserve the structural information between the visual input $\mathbf{x}$ and the generated image $\hat{\mathbf{y}}$.

4.2 Multi-modal Video Prediction

This section is still in progress. We are in the process of conducting more experiments and ablation studies. Explanations and results will be completed very soon.

5 Experiments

5.1 Tactile-Guided Image Stylization

5.1.1 Experimental Setup

Implementation Details Our image encoder and decoder of the generator are fully convolutional neural network consisting of 9 blocks of ResNet-based CNN bottlenecks. The first convolution layer is set to 7×7 and the rest are set to 3×3. For the tactile encoder, we adopt a ResNet-18 [19] backbone pretrained on the ImageNet [12]. For the discriminator we adopt the PatchGAN architecture [20]. To compute the NCE loss, we extract features from five different layers: the input image layer, the first and second downsampling convolution layer and the first and fifth residual blocks. We set the hyperparameter λ and μ equal to 0.5. We train our model on 4 Nvidia 2080-Ti GPUs for 100 epochs with the batch size of 8 and the learning rate of 0.0002. For input visual images, we employ a random crop and an horizontal flip.

5.1.2 Results

We show the qualitative results in the Figure 5. All of the results are generated from the single model (i.e., by one-to-many relation). With input of tactile signals, our model is capable to distinguish and capture the subtle distinction between the input category and the output category without any label.

5.2 Multi-modal Video Prediction

This section is still in progress. We are in the process of conducting more experiments and ablation studies. Explanations and results will be completed very soon.

6 Conclusion

We introduce TOUCH and GO, a multisensory dataset containing real-world synchronized high-quality video and tactile data captured from egocentric viewpoint. Compared to existing real-world multisensory datasets, our work contains much greater hours of videos, object instances, and touches. We propose two novel applications including tactile-guided image stylization and multi-modal video prediction. Leveraging TOUCH and GO dataset, experimental results indicate our designed models outperform label-based counterparts in both quantitative and qualitative evaluations. We hope our dataset, which is easily accessible to the community, will drive more multisensory applications and serve as a standard benchmark.

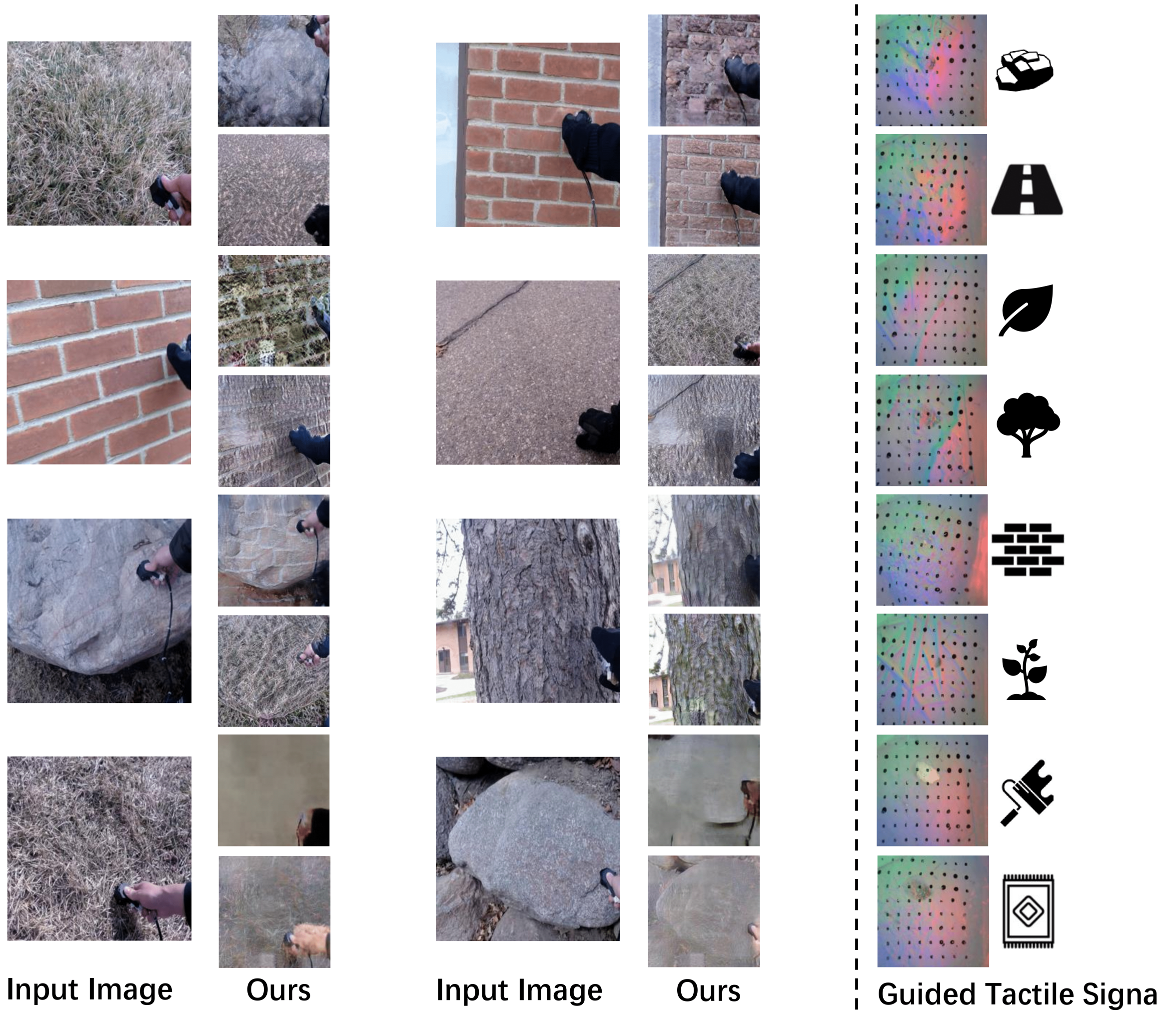

Figure 5: Qualitative results our model on tactile-guided image stylization. For reference, we show guided tactile signals as well as their corresponding images in the last column.

References

[1] *ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models*, Vancouver, Canada, 11/2019 2019.

[2] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H. Campbell, and Sergey Levine. Stochastic variational video prediction. In *International Conference on Learning Representations*, 2018.

[3] Renée Baillargeon. *The Acquisition of Physical Knowledge in Infancy: A Summary in Eight Lessons*, pages 46–83. 11 2002.

[4] Konstantinos Bousmalis and Sergey Levine. Closing the simulation-to-reality gap for deep robotic learning, 2017.

[5] Wonmin Byeon, Qin Wang, Rupesh Kumar Srivastava, and Petros Koumoutsakos. Contextvp: Fully context-aware video prediction. In *ECCV*, 2018.

[6] 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.

[7] Roberto Calandra, Andrew Owens, Manu Upadhyaya, Wenzhen Yuan, Justin Lin, Edward Adelson, and Sergey Levine. The feeling of success: Does touch sensing help predict grasp outcomes? 10 2017.

[8] Berk Calli, Arjun Singh, Aaron Walsman, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M. Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. In *2015 International Conference on Advanced Robotics (ICAR)*, pages 510–517, 2015.

[9] Angel Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository. 12 2015.

[10] Jasmine Collins, Shubham Goel, Achleshwar Luthra, Leon Xu, Kenan Deng, Xi Zhang, Tomas F Yago Vicente, Himanshu Arora, Thomas Dideriksen, Matthieu Guillaumin, and Jitendra Malik. Abo: Dataset and benchmarks for real-world 3d object understanding. *arXiv preprint arXiv:2110.06199*, 2021.

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

[12] 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.

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

[14] Zihan Ding, Nathan Lepora, and Edward Johns. Sim-to-real transfer for optical tactile sensing. pages 1639–1645, 05 2020.

[15] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: A benchmark. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 304–311, 2009.

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

[21] Dinesh Jayaraman, Frederik Ebert, Alexei A. Efros, and Sergey Levine. Time-agnostic prediction: Predicting predictable video frames. *ArXiv*, abs/1808.07784, 2019.

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

[23] 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, 2009.

[24] Nal Kalchbrenner, Aäron van den Oord, Karen Simonyan, Ivo Danihelka, Oriol Vinyals, Alex Graves, and Koray Kavukcuoglu. Video pixel networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1771–1779. PMLR, 06–11 Aug 2017.

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

[26] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *ICML*, 2017.

[27] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2014.

[28] Raj Kolamuri, Zilin Si, Yufan Zhang, Arpit Agarwal, and Wenzhen Yuan. Improving grasp stability with rotation measurement from tactile sensing. In *Proceedings of (IROS) IEEE/RSJ International Conference on Intelligent Robots and Systems*, October 2021.

[29] Oliver Kroemer, Scott Niekum, and George Dimitri Konidaris. A review of robot learning for manipulation: Challenges, representations, and algorithms. *J. Mach. Learn. Res.*, 22:30:1–30:82, 2021.

[30] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020.

[31] S Lederman and Roberta Klatzky. Haptic perception: A tutorial. volume 71, pages 1439–59, 10 2009.

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

[33] 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.

[34] Joseph J. Lim, Hamed Pirsiavash, and Antonio Torralba. Parsing ikea objects: Fine pose estimation. In *2013 IEEE International Conference on Computer Vision*, pages 2992–2999, 2013.

[35] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing.

[36] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 700–708, Red Hook, NY, USA, 2017. Curran Associates Inc.

[37] William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. 05 2016.

[38] Shan Luo, Wenzhen Yuan, Edward Adelson, Anthony Cohn, and Raul Fuentes. Cloth texture recognition using vision and tactile sensing. 05 2018.

[39] Michaël Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *CoRR*, abs/1511.05440, 2016.

[40] Vincent Michalski, Roland Memisevic, and Kishore Konda. Modeling deep temporal dependencies with recurrent grammar cells"". 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.

[41] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets, 2014.

[42] Roni Mittelman, Benjamin Kuipers, Silvio Savarese, and Honglak Lee. Structured recurrent temporal restricted boltzmann machines. In *ICML*, pages 1647–1655, 2014.

[43] 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.

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

[45] Patsorn Sangkloy, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. Scribbler: Controlling deep image synthesis with sketch and color. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6836–6845, 2017.

[46] Laura Schulz. The origins of inquiry: Inductive inference and exploration in early childhood. *Trends in cognitive sciences*, 16:382–9, 06 2012.

[47] Arjun Singh, James Sha, Karthik S. Narayan, Tudor Achim, and P. Abbeel. Bigbird: A large-scale 3d database of object instances. *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 509–516, 2014.

[48] Linda B. Smith and Michael Gasser. The development of embodied cognition: Six lessons from babies. *Artificial Life*, 11:13–29, 2005.

[49] Elizabeth S. Spelke and Katherine D. Kinzler. Core knowledge. *Developmental science*, 10 1:89–96, 2007.

[50] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhutdinov. Unsupervised learning of video representations using LSTMs. In *ICML*, 2015.

[51] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[52] Ilya Sutskever, Geoffrey E. Hinton, and Graham W. Taylor. The recurrent temporal restricted boltzmann machine. In *NIPS*, 2008.

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

[54] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. NIPS'18, page 1152–1164, Red Hook, NY, USA, 2018. Curran Associates Inc.

[55] 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. pages 8798–8807, 06 2018.

[56] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. pages 1912–1920, 06 2015.

[57] Yu Xiang, Wonhui Kim, Wei Chen, Jingwei Ji, Chris Choy, Hao Su, Roozbeh Mottaghi, Leonidas Guibas, and Silvio Savarese. Objectnet3d: A large scale database for 3d object recognition. volume 9912, pages 160–176, 10 2016.

[58] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2868–2876, 2017.

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

[60] Wenzhen Yuan, Shaoxiong Wang, Siyuan Dong, and Edward Adelson. Connecting look and feel: Associating the visual and tactile properties of physical materials. 04 2017.

[61] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei Efros. Generative visual manipulation on the natural image manifold. volume 9909, 08 2016.

[62] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. pages 2242–2251, 10 2017.

[63] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 465–476. Curran Associates, Inc., 2017.