Connecting Touch and Vision via Cross-Modal Prediction

Yunzhu Li Jun-Yan Zhu Russ Tedrake Antonio Torralba
MIT CSAIL

Abstract

Humans perceive the world using multi-modal sensory inputs such as vision, audition, and touch. In this work, we investigate the cross-modal connection between vision and touch. The main challenge in this cross-domain modeling task lies in the significant scale discrepancy between the two: while our eyes perceive an entire visual scene at once, humans can only feel a small region of an object at any given moment. To connect vision and touch, we introduce new tasks of synthesizing plausible tactile signals from visual inputs as well as imagining how we interact with objects given tactile data as input. To accomplish our goals, we first equip robots with both visual and tactile sensors and collect a large-scale dataset of corresponding vision and tactile image sequences. To close the scale gap, we present a new conditional adversarial model that incorporates the scale and location information of the touch. Human perceptual studies demonstrate that our model can produce realistic visual images from tactile data and vice versa. Finally, we present both qualitative and quantitative experimental results regarding different system designs, as well as visualizing the learned representations of our model.

1. Introduction

People perceive the world in a multi-modal way where vision and touch are highly intertwined [24, 42]: when we close our eyes and use only our fingertips to sense an object in front of us, we can make guesses about its texture and geometry. For example in Figure 1d, one can probably tell that s/he is touching a piece of delicate fabric based on its tactile "feeling"; similarly, we can imagine the feeling of touch by just seeing the object. In Figure 1c, without directly contacting the rim of a mug, we can easily imagine the sharpness and hardness of the touch merely by our visual perception. The underlying reason for this cross-modal connection is the shared physical properties that influence both modalities such as local geometry, texture, roughness, hardness and so on. Therefore, it would be desirable to build a computational model that can extract such shared representations from one modality and transfer them to the other.

In this work, we present a cross-modal prediction system

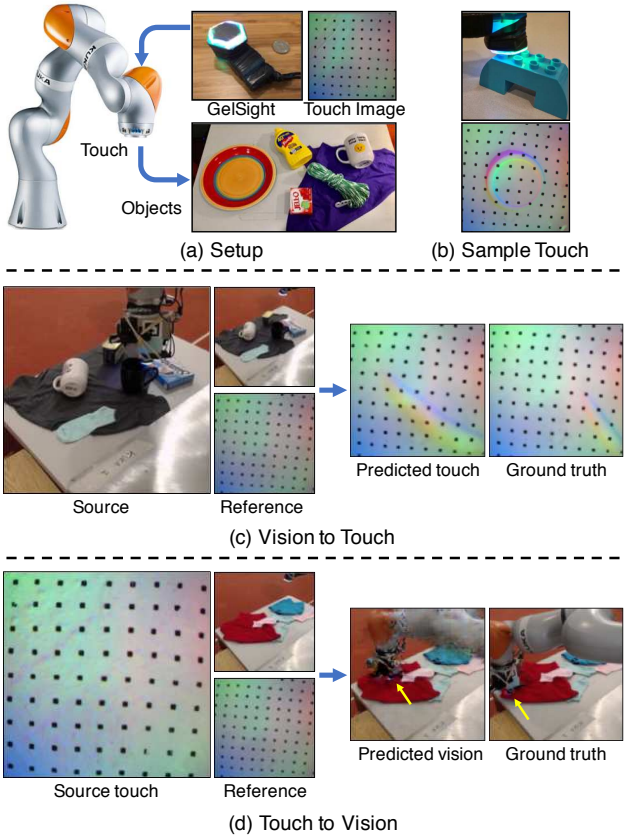

Figure 1. **Data collection setup:** (a) we use a robot arm equipped with a GelSight sensor [15] to collect tactile data and use a webcam to capture the videos of object interaction scenes. (b) An illustration of the GelSight touching an object. **Cross-modal prediction:** given the collected vision-tactile pairs, we train cross-modal prediction networks for several tasks: (c) Learning to feel by seeing (vision → touch): predicting the a touch signal from its corresponding vision input and reference images and (d) Learning to see by touching (touch → vision): predicting vision from touch. The predicted touch locations and ground truth locations (marked with yellow arrows in (d)) share a similar feeling. Please check out our website for code and more results.

between vision and touch with the goals of *learning to see by touching* and *learning to feel by seeing*. Different from other cross-modal prediction problems where sensory data in different domains are roughly spatially aligned [13, 1], the

scale gap between vision and touch signals is huge. While our visual perception system processes the entire scene as a whole, our fingers can only sense a tiny fraction of the object at any given moment. To investigate the connections between vision and touch, we introduce two cross-modal prediction tasks: (1) synthesizing plausible temporal tactile signals from vision inputs, and (2) predicting which object and which object part is being touched directly from tactile inputs. Figure 1c and d show a few representative results.

To accomplish these tasks, we build a robotic system to automate the process of collecting large-scale visual-touch pairs. As shown in Figure 1a, a robot arm is equipped with a tactile sensor called GelSight [15]. We also set up a standalone web camera to record visual information of both objects and arms. In total, we recorded $12,000$ touches on 195 objects from a wide range of categories. Each touch action contains a video sequence of 250 frames, resulting in 3 million visual and tactile paired images. The usage of the dataset is not limited to the above two applications.

Our model is built on conditional adversarial networks [11, 13]. The standard approach [13] yields less satisfactory results in our tasks due to the following two challenges. First, the scale gap between vision and touch makes the previous methods [13, 39] less suitable as they are tailored for spatially aligned image pairs. To address this scale gap, we incorporate the scale and location information of the touch into our model, which significantly improves the results. Second, we encounter severe mode collapse during GANs training when the model generates the same output regardless of inputs. It is because the majority of our tactile data only contain flat regions as often times, robots arms are either in the air or touching textureless surface, To prevent mode collapse, we adopt a data rebalancing strategy to help the generator produce diverse modes.

We present both qualitative results and quantitative analysis to evaluate our model. The evaluations include human perceptual studies regarding the photorealism of the results, as well as objective measures such as the accuracy of touch locations and the amount of deformation in the GelSight images. We also perform ablation studies regarding alternative model choices and objective functions. Finally, we visualize the learned representations of our model to help understand what it has captured.

2. Related Work

Cross-modal learning and prediction People understand our visual world through many different modalities. Inspired by this, many researchers proposed to learn shared embeddings from multiple domains such as words and images [9], audio and videos [32, 2, 36], and texts and visual data [33, 34, 1]. Our work is mostly related to cross-modal prediction, which aims to predict data in one domain from another. Recent work has tackle different prediction tasks such as using vision to predict sound [35] and generating captions for images [20, 17, 41, 6], thanks to large-scale paired cross-domain datasets, which are not currently available for vision and touch. We circumvent this difficulty by automating the data collection process with robots.

Vision and touch To give intelligent robots the same tactile sensing ability, different types of force, haptic, and tactile sensors [22, 23, 5, 16] have been developed over the decades. Among them, GelSight [14, 15, 43] is considered among the best high-resolution tactile sensors. Recently, researchers have used GelSight and other types of force and tactile sensors for many vision and robotic applications [45, 47, 44, 26, 27, 25]. Yuan et al. [46] studied physical and material properties of fabrics by fusing visual, depth, and tactile sensors. Calandra et al. [3] proposed a visual-tactile model for predicting grasp outcomes. Different from prior work that used vision and touch to improve individual tasks, in this work we focus on several cross-modal prediction tasks, investigating whether we can predict one signal from the other.

Image-to-image translation Our model is built upon recent work on image-to-image translation [13, 28, 50], which aims to translate an input image from one domain to a photorealistic output in the target domain. The key to its success relies on adversarial training [11, 30], where a discriminator is trained to distinguish between the generated results and real images from the target domain. This method enables many applications such as synthesizing photos from user sketches [13, 38], changing night to day [13, 51], and turning semantic layouts into natural scenes [13, 39]. Prior work often assumes that input and output images are geometrically aligned, which does not hold in our tasks due to the dramatic scale difference between two modalities. Therefore, we design objective functions and architectures to sidestep this scale mismatch. In Section 5, we show that we can obtain more visually appealing results compared to recent methods [13].

3. VisGel Dataset

Here we describe our data collection procedure including the tactile sensor we used, the way that robotic arms interacting with objects, and a diverse object set that includes 195 different everyday items from a wide range of categories.

Data collection setup Figure 1a illustrates the setup in our experiments. We use KUKA LBR iiwa industrial robotic arms to automate the data collection process. The arm is equipped with a GelSight sensor [43] to collect raw tactile images. We set up a webcam on a tripod at the back of the arm to capture videos of the scenes where the robotic arm touching the objects. We use recorded timestamps to synchronize visual and tactile images.

(a) Training objects and known objects for testing (b) Unseen test objects

Figure 2. **Object set.** Here we show the object set used in training and test. The dataset includes a wide range of objects from food items, tools, kitchen items, to fabrics and stationery.

GelSight sensor The GelSight sensor [14, 15, 43] is an optical tactile sensor that measures the texture and geometry of a contact surface at very high spatial resolution [15]. The surface of the sensor is a soft elastomer painted with a reflective membrane that deforms to the shape of the object upon contact, and the sensing area is about 1.5cm $\times$ 1.5cm. Underneath this elastomer is an ordinary camera that views the deformed gel. The colored LEDs illuminate the gel from different directions, resulting in a three-channel surface normal image (Figure 1b). GelSight also uses markers on the membrane and recording the flow field of the marker movement to sketched the deformation. The 2D image format of the raw tactile data allows us to use standard convolutional neural networks (CNN) [21] for processing and extracting tactile information. Figure 1c and d show a few examples of collected raw tactile data.

Objects dataset Figure 2 shows all the 195 objects used in our study. To collect such a diverse set of objects, we start from Yale-CMU-Berkeley (YCB) dataset [4], a standard daily life object dataset widely used in robotic manipulation research. We use 45 objects with a wide range of shapes, textures, weight, sizes, and rigidity. We discard the rest of the 25 small objects (e.g., plastic nut) as they are occluded by the robot arm from the camera viewpoint. To further increase the diversity of objects, we obtain additional 150 new consumer products that include the categories in the YCB dataset (i.e., food items, tool items, shape items, task items, and kitchen items) as well as new categories such as fabrics and stationery. We use 165 objects during our training and 30 seen and 30 novel objects during test. Each scene contains $4 \sim 10$ randomly placed objects that sometimes overlap with each other.

Generating touch proposals A random touch at an arbitrary location may be suboptimal due to two reasons. First, the robotic arm can often touch nothing but the desk. Second, the arm may touch in an undesirable direction or unexpectedly move the object so that the GelSight Sensor fails to capture any tactile signal. To address the above issues and generate better touch proposals, we first reconstruct 3D point clouds of the scene with a real-time SLAM system called ElasticFusion [40]. We then sample a random touch region whose surface normals are mostly perpendicular to the desk. The touching direction is important as it allows robot arms to firmly press the object without moving it.

	# touches	# total vision-touch frames
Train	10,000	2,500,000
Test	2,000	500,000

Table 1. **Statistics of our VisGel dataset.** We use a video camera and a tactile sensor to collect a large-scale synchronized videos of a robot arm interacting with household objects.

Dataset Statistics We have collected synchronized tactile images and RGB images for 195 objects. Table 1 shows the basic statistics of the dataset for both training and test. To our knowledge, this is the largest vision-touch dataset.

4. Cross-Modal Prediction

We propose a cross-modal prediction method for predicting vision from touch and vice versa. First, we describe our basic method based on conditional GANs [13] in Section 4.1. We further improve the accuracy and the quality of our prediction results with three modifications tailored for our tasks in Section 4.2. We first incorporate the scale and location of the touch into our model. Then, we use a data rebalancing mechanism to increase the diversity of our results. Finally, we further improve the temporal coherence and accuracy of our results by extracting temporal information from nearby input frames. In Section 4.3, we describe the details of our training procedure as well as network designs.

4.1. Conditional GANs

Our approach is built on the pix2pix method [13], a recently proposed general-purpose conditional GANs framework for image-to-image translation. In the context of vision-touch cross-modal prediction, the generator G takes either a vision or tactile image x as an input and produce an output image in the other domain with $y = G(x)$. The discriminator D observes both the input image x and the output result y: $D(x, y) \rightarrow [0, 1]$. During training, the discriminator D is trained to reveal the differences between synthesized results and real images while the objective of the generator G is to produce photorealistic results that can fool the discriminator D. We train the model with vision-touch image pairs $\{(\mathbf{x}, \mathbf{y})\}$. In the task of touch $\rightarrow$ vision, $\mathbf{x}$ is a touch image and $\mathbf{y}$ is the corresponding visual image. The same thing applies to the vision $\rightarrow$ touch direction, i.e., $(\mathbf{x}, \mathbf{y}) = (\text{visual image}, \text{touch image})$. Conditional GANs can be optimized via the following min-max objective:

$$G^* = \arg\min_G \max_D \mathcal{L}_{\text{GAN}}(G, D) + \lambda \mathcal{L}_1(G) \qquad (1)$$

where the adversarial loss $\mathcal{L}_{\text{GAN}}(G, D)$ is derived as:

$$\mathbb{E}_{(\mathbf{x},\mathbf{y})}[\log D(\mathbf{x}, \mathbf{y})] + \mathbb{E}_{\mathbf{x}}[\log(1 - D(\mathbf{x}, G(\mathbf{x}))], \qquad (2)$$

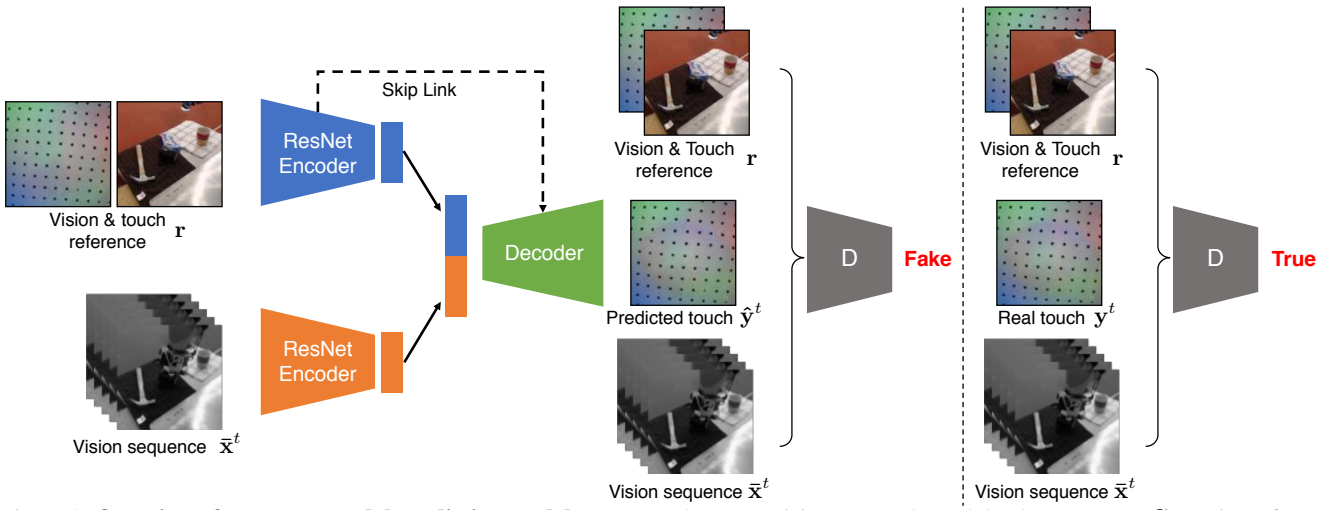

Figure 3. **Overview of our cross-modal prediction model.** Here we show our vision → touch model. The generator G consists of two ResNet encoders and one decoder. It takes both reference vision and touch images $\mathbf{r}$ as well as a sequence of frames $\bar{\mathbf{x}}^t$ as input, and predict the tactile signal $\hat{\mathbf{y}}^t$ as output. Both reference images and temporal information help improve the results. Our discriminator learns to distinguish between the generated tactile signal $\hat{\mathbf{y}}^t$ and real tactile data $\mathbf{y}^t$. For touch → vision, we switch the input and output modality and train the model under the same framework.

where the generator G strives to minimize the above objective against the discriminator's effort to maximize it, and we denote $\mathbb{E}_{\mathbf{x}} \triangleq \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(x)}$ and $\mathbb{E}_{(\mathbf{x},\mathbf{y})} \triangleq \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p_{\text{data}}(\mathbf{x},\mathbf{y})}$ for brevity. In additional to the GAN loss, we also add a direct regression L_1 loss between the predicted results and the ground truth images. This loss has been shown to help stabilize GAN training in prior work [13]:

$$\mathcal{L}_1(G) = \mathbb{E}_{(\mathbf{x},\mathbf{y})}||\mathbf{y} - G(\mathbf{x})||_1 \tag{3}$$

4.2. Improving Photorealism and Accuracy

We first experimented with the above conditional GANs framework. Unfortunately, as shown in the Figure 4, the synthesized results are far from satisfactory, often looking unrealistic and suffering from severe visual artifacts. Besides, the generated results do not align well with input signals.

To address the above issues, we make a few modifications to the basic algorithm, which significantly improve the quality of the results as well as the match between input-output pairs. We first feed tactile and visual reference images to both the generator and the discriminator so that the model only needs to learn to model cross-modal changes rather than the entire signal. Second, we use a data-driven data rebalancing mechanism in our training so that the network is more robust to mode collapse problem where the data is highly imbalanced. Finally, we extract information from multiple neighbor frames of input videos rather than the current frame alone, producing temporal coherent outputs.

Using reference tactile and visual images As we have mentioned before, the scale between a touch signal and a visual image is huge as a GelSight sensor can only contact a very tiny portion w.r.t. the visual image. This makes the cross-modal prediction between vision and touch quite challenging. Regarding touch to vision, we need to solve an almost impossible 'extrapolation' problem from a tiny patch to an entire image. From vision to touch, the model has to first localize the location of the touch and then infer the material and geometry of the touched region. Figure 4 shows a few results produced by conditional GANs model described in Section 4.1, where no reference is used. The low quality of the results is not surprising due to self-occlusion and big scale discrepancy.

We sidestep this difficulty by providing our system both the reference tactile and visual images as shown in Figure 1c and d. A reference visual image captures the original scene without any robot-object interaction. For vision to touch direction, when the robot arm is operating, our model can simply compare the current frame with its reference image and easily identify the location and the scale of the touch. For touch to vision direction, a reference visual image can tell our model the original scene and our model only needs to predict the location of the touch and hallucinate the robotic arm, without rendering the entire scene from scratch. A reference tactile image captures the tactile response when the sensor touches nothing, which can help the system calibrate the tactile input, as different GelSight sensors have different lighting distribution and black dot patterns.

In particular, we feed both vision and tactile reference images $\mathbf{r} = (\mathbf{x}^{\text{ref}}, \mathbf{y}^{\text{ref}})$ to the generator G and the discriminator D. As the reference image and the output often share common low-level features, we introduce skip connections [37, 12] between the encoder convolutional layers and transposed-convolutional layers in our decoder.

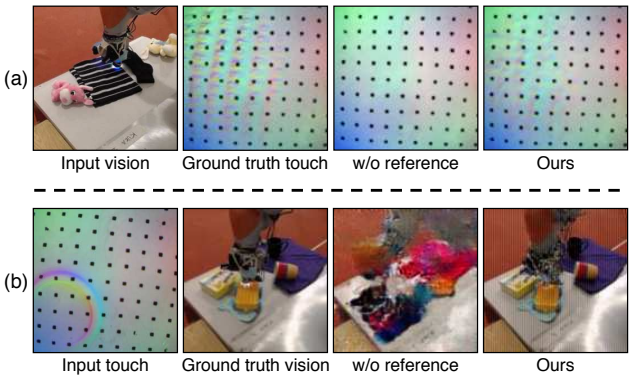

Figure 4. **Using reference images.** Qualitative results of our methods with / without using reference images. Our model trained with reference images produces more visually appealing images.

Data rebalancing In our recorded data, around 60 percentage of times, the robot arm is in the air without touching any object. This results in a huge data imbalance issue, where more than half of our tactile data has only near-flat responses without any texture or geometry. This highly imbalanced dataset causes severe model collapse during GANs training [10]. To address it, we apply data rebalancing technique, widely used in classification tasks [8, 48]. In particular, during the training, we reweight the loss of each data pair $(\mathbf{x}^t, \mathbf{r}, \mathbf{y}^t)$ based on its rarity score $\mathbf{w}^t$. In practice, we calculate the rarity score based on a ad-hoc metric. We first compute a residual image $\|\mathbf{x}^t - \mathbf{x}^{\text{ref}}\|$ between the current tactile data $\mathbf{x}^t$ and its reference tactile data $\mathbf{x}^{\text{ref}}$. We then simply calculate the variance of Laplacian derivatives over the difference image. For IO efficiency, instead of reweighting, we sample the training data pair $(\mathbf{x}^t, \mathbf{r}, \mathbf{y}^t)$ with the probability $\frac{\mathbf{w}^t}{\sum_t \mathbf{w}^t}$. We denote the resulting data distribution as $p_{\mathbf{w}}$. Figure 5 shows a few qualitative results demonstrating the improvement by using data rebalancing. Our evaluation in Section 5 also shows the effectiveness of data rebalancing.

Incorporating temporal cues We find that our initial results look quite realistic, but the predicted output sequences and input sequences are often out of sync (Figure 7). To address this temporal mismatch issue, we use multiple nearby frames of the input signal in addition to its current frame. In practice, we sample 5 consecutive frames every 2 frames: $\bar{\mathbf{x}}^t = \{\mathbf{x}^{t-4}, \mathbf{x}^{t-2}, \mathbf{x}^t, \mathbf{x}^{t+2}, \mathbf{x}^{t+4}\}$ at a particular moment t. To reduce data redundancy, we only use grayscale images and leave the reference image as RGB.

Our full model Figure 3 shows an overview of our final cross-modal prediction model. The generator G takes both input data $\bar{\mathbf{x}}^t = \{\mathbf{x}^{t-4}, \mathbf{x}^{t-2}, \mathbf{x}^t, \mathbf{x}^{t+2}, \mathbf{x}^{t+4}\}$ as well as reference vision and tactile images $\mathbf{r} = (\mathbf{x}^{\text{ref}}, \mathbf{y}^{\text{ref}})$ and produce a output image $\hat{\mathbf{y}}^t = G(\bar{\mathbf{x}}^t, \mathbf{r})$ at moment t in the target domain. We extend the minimax objective (Equation 1) $\mathcal{L}_{\text{GAN}}(G, D) + \lambda \mathcal{L}_1(G)$, where $\mathcal{L}_{\text{GAN}}(G, D)$ is as follows:

$$\mathbb{E}_{(\bar{\mathbf{x}}^t, \mathbf{r}, \mathbf{y}^t)}[\log D(\bar{\mathbf{x}}^t, \mathbf{r}, \mathbf{y}^t)] + \mathbb{E}_{(\bar{\mathbf{x}}^t, \mathbf{r})}[\log(1 - D(\bar{\mathbf{x}}^t, \mathbf{r}, \hat{\mathbf{y}}^t)], \tag{4}$$

where G and D both takes both temporal data $\bar{\mathbf{x}}^t$ and reference images $\mathbf{r}$ as inputs. Similarly, the regression loss $\mathcal{L}_1(G)$ can be calculated as:

$$\mathcal{L}_1(G) = \mathbb{E}_{(\bar{\mathbf{x}}^t, \mathbf{r}, \mathbf{y}^t) \sim p_{\mathbf{w}}} ||\mathbf{y}^t - \hat{\mathbf{y}}^t||_1 \tag{5}$$

Figure 3 shows a sample input-output combination where the network takes a sequence of visual images and the corresponding references as inputs, synthesizing a tactile prediction as output. The same framework can be applied to the touch $\rightarrow$ vision direction as well.

4.3. Implementation details

Network architectures We use an encoder-decoder architecture for our generator. For the encoder, we use two ResNet18 models [12] for encoding input images $\mathbf{x}$ and reference tactile and visual images $\mathbf{r}$ into 512 dimensional latent vectors respectively. We concatenate two vectors from both encoders into one 1024 dimensional vector and feed it to the decoder that contains 5 standard strided-convolution layers. As the output result looks close to one of the reference images, we add a few skip connections between the convolutional layers from the reference branch in our encoder and strided-convolutional layers in the decoder. For the discriminator, we use a standard ConvNets with multiple convolutional layers. Please find more architectural details in our supplementary material.

Training We train the models with the Adam solver [18] with a learning rate of 0.0002. We set $\lambda = 10$ for $\mathcal{L}_1$ loss. We use LSGANs loss [29] rather than standard GANs [11] for more stable training, as shown in prior work [50, 39]. We apply standard data augmentation techniques [19] including random cropping and slightly perturbing the brightness, contrast, saturation, and hue of input images.

5. Experiments

We evaluate our method on cross-modal prediction tasks between vision and touch using the VisGel dataset. We report multiple metrics that evaluate different aspects of the predictions. When predicting touch from vision, we measure (1) perceptual realism using AMT: whether results look realistic, (2) the moment of contact: whether our model can predict if a GelSight sensor is in contact with the object, and (3) markers' deformation: whether our model can track the deformation of the membrane. Regarding touch $\rightarrow$ vision direction, we evaluate our model using (1) visual realism via AMT and (2) the sense of touch: whether the predicted touch position shares a similar feel with the ground truth position. We also include the evaluations regarding full-reference metrics in the supplement. Please find our code, data, and more results on our website.

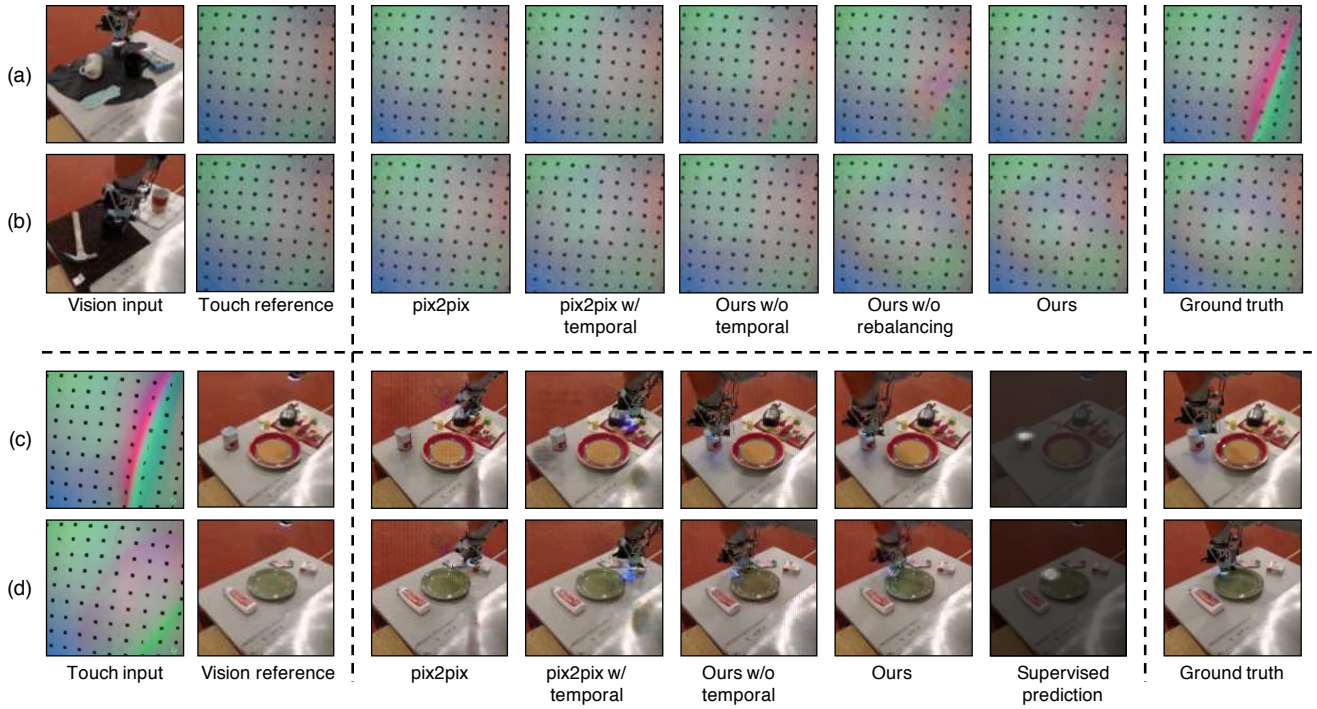

Figure 5. **Example cross-modal prediction results.** (a) and (b) show two examples of vision → touch prediction by our model and baselines. (c) and (d) show the touch → vision direction. In both cases, our results appear both realistic and visually similar to the ground truth target images. In (c) and (d), our model, trained without ground truth position annotation, can accurately predict touch locations, comparable to a fully supervised prediction method.

Method	Seen objects % Turkers labeled *real*	Unseen objects % Turkers labeled *real*
pix2pix [13]	28.09 %	21.74%
pix2pix [13] w/ temporal	35.02%	27.70%
Ours w/o temporal	41.44%	31.60%
Ours w/o rebalancing	39.95%	34.86%
Ours	**46.63%**	**38.22%**

Table 2. **Vision2Touch AMT "real vs fake" test.** Our method can synthesize more realistic tactile signals, compared to both pix2pix [13] and our baselines, both for seen and novel objects.

We feed reference images to all the baselines, as they are crucial for handling the scale discrepancy (Figure 4).

5.1. Vision → Touch

In this experiment, we compare our method with various baselines. We first run the trained models to generate GelSight outputs frame by frame and then concatenate adjacent frames together into a video. Each video contains exactly one action with 64 consecutive frames.

An ideal model should produce a perceptually realistic and temporal coherent output. Furthermore, when humans observe this kind of physical interaction, we can roughly infer the moment of contact and the force being applied to the touch; hence, we would also like to assess our model's understanding of the interaction. In particular, we evaluate whether our model can predict the moment of contact as well as the deformation of the markers grid.

Perceptual realism (AMT) Human subjective ratings have been shown to be a more meaningful metric for image synthesis tasks [48, 13] compared to metrics such as RMS or SSIM. We follow a similar study protocol as described in Zhang et al. [48] and run a real vs. fake forced-choice test on Amazon Mechanical Turk (AMT). In particular, we present our participants with the ground truth tactile videos and the predicted tactile results along with the vision inputs. We ask which tactile video corresponds better to the input vision signal. As most people may not be familiar with tactile data, we first educate the participants with 5 typical ground truth vision-touch video pairs and detailed instruction. In total, we collect $8,000$ judgments for $1,250$ results. Table 2 shows that our full method can outperform the pix2pix [13] as well as the alternative design choices on both seen objects (different touches) and unseen objects.

The moment of contact The deformation on the GelSight marker field indicates whether and when a GelSight sensor touches the surface. We evaluate our system by measuring how well it can predict the *moment of contact* with respect to the ground truth data from tactile sensors. We track the GelSight markers and calculate the average L_2 distance for each marker. For each touch episode, we denote the largest deformation distance as d_{max} and the smallest deformation as d_{min}, we then set a cutoff threshold at $r \cdot (d_{\text{max}} - d_{\text{min}}) + d_{\text{min}}$, where r is set to 0.6 in our case. We mark the left most and right most cutoff time point as t_l and t_r individually.

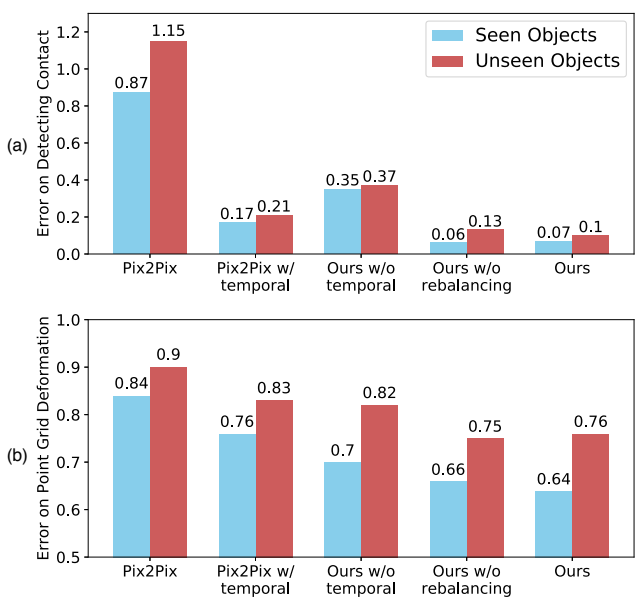

Figure 6. **Vision2Touch - quantitative results. Top:** Errors on detecting the moment of contact. Our method generally performs the best. The use of temporal cues can significantly improve the performance or our model. **Bottom:** Errors on the average markers' deformation. Our method still works best.

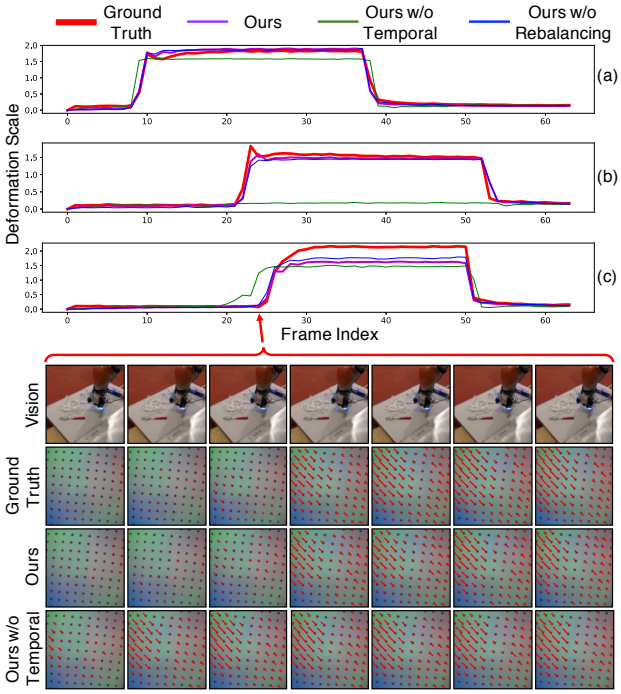

Figure 7. **Vision2Touch - detecting the moment of contact.** We show the markers' deformation across time, determined by the average shift of all black markers. Higher deformation implies object contact with a larger force. **Top:** Three typical cases, where (a) all methods can infer the moment of contact, (b) the method without temporal cues failed to capture the moment of contact, and (c) the method without temporal cues produces misaligned results. **Bottom:** We show several vision and touch frames from case (c). Our model with temporal cues can predict GelSight's deformation more accurately. The motion of the markers is magnified in red for better visualization.

Similarly, we compute the ground truth cutoff time as t_l^{gt} and t_r^{gt}; then the error of the *moment of contact* for this episode is determined as $e_{\text{contact}} = |t_l - t_l^{\text{gt}}| + |t_r - t_r^{\text{gt}}|$.

As shown in Figure 6a, the methods without temporal cues produce a large error due to temporal misalignment. Meanwhile, our model works better on seen objects than unseen objects, which coincides with the empirical observation that human beings can better predict the touch outcomes if we have interacted with the object before.

We also show a few deformation curves over the time. Figure 7a illustrates a case where all the methods perform well in detecting the ground truth *moment of contact*. Figure 7b shows an example where the model without temporal cues completely missed the contact event. Figure 7c shows another common situation, where the *moment of contact* predicted by the method without temporal cues shifts from the ground truth. Figure 7 shows several ground truth and the predicted frames. As expected, a single-frame method fails to accurately predict the contact moment.

Tracking markers' deformation The flow field of the GelSight markers characterizes the deformation of the membrane, which is useful for representing contact forces as well as detecting slippery [7]. In this section, we assess our model's ability by comparing the predicted deformation with the ground truth deformation. We calculate the average L_2 distance between each corresponding markers in the ground truth and the generated touch image. Figure 6b shows that the single-frame model performs the worst, as it misses important temporal cues, which makes it hard to infer information like force and sliding.

Visualizing the learned representation We visualize the learned representation using a recent network interpretation method [49], which can highlight important image regions for final decisions (Figure 8a and b). Many meaningful patterns emerge, such as arms hovering in the air or touching flat area and sharp edges. This result implies that our representation learns shared information across two modalities. Please see our supplemental material for more visualizations.

5.2. Touch → Vision

We can also go from touch to vision - by giving the model a reference visual image and tactile signal, can the model imagine where it is touching? It is impossible to locate the GelSight if the sensor is not in contact with anything; hence, we only include vision-touch pairs where the sensor touches the objects. The model shall predict a reasonable touch position based on the geometric cues from the touch images.

The Sense of Touch Different regions and objects can stimulate similar senses of touch. For example, our finger

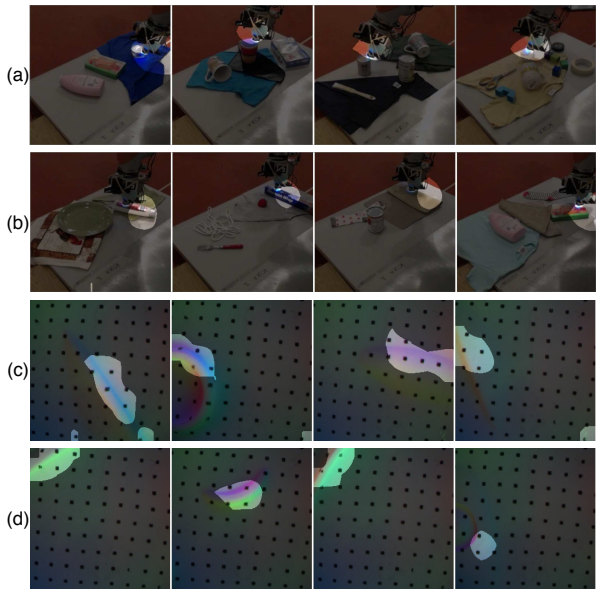

Figure 8. **Visualizing the learned representations using Zhou et al. [49]** (a) and (b) visualize two internal units of our vision → touch model. They both highlight the position of the GelSight, but focus on sharp edges and flat surfaces respectively. (c) and (d) visualize internal units of our touch → vision model. They focus on different geometric patterns.

may feel the same if we touch various regions on a flat surface or along the same sharp edge. Therefore, given a tactile input, it is unrealistic to ask a model to predict the exact same touch location as the ground truth. As long as the model can predict a touch position that feels the same as the ground truth position, it is still a valid prediction. To quantify this, we show the predicted visual image as well as the ground truth visual image. Then we ask human participants whether these two touch locations feel similar or not. We report average accuracies over 400 images per method (200 known objects, 200 unknown objects), and Table 3 shows the performance of different methods. Our full method can produce much more plausible touch positions.

We also compare our method with a baseline trained with external supervisions provided by humans. Specifically, we hand-label the position of the GelSight on 1,000 images, and train a Stacked Hourglass Networks [31] to predict possible touch positions. Qualitative and quantitative comparisons are shown in Figure 5 and Table 3. Our self-supervised method is comparable to its fully-supervised counterpart.

Perceptual Realism (AMT) Since it is difficult for humans participants to imagine a robotic manipulation scene given only a single tactile data, we only evaluate the quality of results without showing the tactile input. In particular, we show each image for 1 second, and AMT participants are then given unlimited time to decide which one is fake. The first 10 images of each HITs are used for practice and we give AMT participants the correct answer. The participants

	Method	Seen objects % Turkers labeled *Feels Similar*	Unseen objects % Turkers labeled *Feels Similar*
Self-Supervised	pix2pix [13]	44.52%	25.21%
	pix2pix w/ temporal	53.27%	35.45%
	Ours w/o temporal	81.31%	78.40%
	Ours	**89.20%**	**83.44%**
	Supervised prediction	**90.37%**	**85.29%**

Table 3. **Touch2Vision "Feels Similar vs Feels Different" test**. Our self-supervised method significantly outperforms baselines. The accuracy is comparable to fully supervised prediction method trained with ground truth annotations.

Method	Seen objects % Turkers labeled *real*	Unseen objects % Turkers labeled *real*
pix2pix [13]	25.80%	**26.13%**
pix2pix [13] w/ temporal	23.61%	19.67%
Ours w/o temporal	**30.80%**	20.74%
Ours	30.50%	24.22%

Table 4. **Touch2Vision AMT "real vs fake" test**. Although pix2pix achieves the highest score for unseen objects, it always produces identical images due to mode collapse. Figure 5 shows a typical collapsed mode, where pix2pix always places the arm at the top-right corner of the image. More qualitative results can be found in our supplementary material.

are then asked to finish the next 40 trials.

In total, we collect 8,000 judgments for 1,000 results. Table 4 shows the fooling rate of each method. We note that results from pix2pix [13] suffer from severe mode collapse and always produce identical images, although they look realistic according to AMT participants. See our website for a more detailed comparison. We also observe that the temporal cues do not always help improve the quality of results for touch → vision direction as we only consider vision-touch pairs during the moment of touch.

Visualizing the learned representation The visualization of the learned representations (Figure 8c and d) show two units that focus on different geometric cues. Please see our supplemental material for more examples.

6. Discussion

In this work, we have proposed to draw connections between vision and touch with conditional adversarial networks. Humans heavily rely on both sensory modalities when interacting with the world. Our model can produce promising cross-modal prediction results for both known objects and unseen objects. In the future, vision-touch cross-modal connection may help downstream vision and robotics applications, such as object recognition and grasping in a low-light environment, and physical scene understanding.

Acknowledgement

This work was supported by: Draper Laboratory Incorporated, Sponsor Award No. SC001-0000001002; NASA - Johnson Space Center, Sponsor Award No. NNX16AC49A.

References

[1] Yusuf Aytar, Lluis Castrejon, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Cross-modal scene networks. *PAMI*, 2017. 1, 2

[2] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *NIPS*, 2016. 2

[3] 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? In *PMLR*, 2017. 2

[4] Berk Calli, Arjun Singh, James Bruce, Aaron Walsman, Kurt Konolige, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. Yale-cmu-berkeley dataset for robotic manipulation research. *The International Journal of Robotics Research*, 2017. 3

[5] Mark R Cutkosky, Robert D Howe, and William R Provancher. Force and tactile sensors. In *Springer Handbook of Robotics*, 2008. 2

[6] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015. 2

[7] Siyuan Dong, Wenzhen Yuan, and Edward Adelson. Improved gelsight tactile sensor for measuring geometry and slip. In *IROS*, 2017. 7

[8] Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. Learning hierarchical features for scene labeling. *PAMI*, 2013. 5

[9] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013. 2

[10] Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016. 5

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. In *NIPS*, 2014. 2, 5

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4, 5

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 1, 2, 3, 4, 6, 8

[14] Micah K Johnson and Edward H Adelson. Retrographic sensing for the measurement of surface texture and shape. In *CVPR*, 2009. 2, 3

[15] Micah K Johnson, Forrester Cole, Alvin Raj, and Edward H Adelson. Microgeometry capture using an elastomeric sensor. In *SIGGRAPH*, 2011. 1, 2, 3

[16] Zhanat Kappassov, Juan-Antonio Corrales, and Véronique Perdereau. Tactile sensing in dexterous robot hands. *Robotics and Autonomous Systems*, 2015. 2

[17] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 2

[18] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014. 5

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 5

[20] Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. Babytalk: Understanding and generating simple image descriptions. *PAMI*, 2013. 2

[21] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998. 3

[22] Susan J Lederman and Roberta L Klatzky. Hand movements: A window into haptic object recognition. *Cognitive psychology*, 1987. 2

[23] Susan J Lederman and Roberta L Klatzky. Haptic perception: A tutorial. *Attention, Perception, & Psychophysics*, 2009. 2

[24] Susan J Lederman, Georgie Thorne, and Bill Jones. Perception of texture by vision and touch: Multidimensionality and intersensory integration. *Journal of Experimental Psychology: Human Perception and Performance*, 1986. 1

[25] 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. In *ICRA*, 2019. 2

[26] Rui Li and Edward H Adelson. Sensing and recognizing surface textures using a gelsight sensor. In *CVPR*, 2013. 2

[27] Rui Li, Robert Platt, Wenzhen Yuan, Andreas ten Pas, Nathan Roscup, Mandayam A Srinivasan, and Edward Adelson. Localization and manipulation of small parts using gelsight tactile sensing. In *IROS*, 2014. 2

[28] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *NIPS*, 2017. 2

[29] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *ICCV*, 2017. 5

[30] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2

[31] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 8

[32] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *ICML*, 2011. 2

[33] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. In *ICLR*, 2014. 2

[34] Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Naokazu Yokoya. Learning joint representations of videos and sentences with web image search. In *ECCV*, 2016. 2

[35] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. Visually indicated sounds. In *CVPR*, 2016. 2

[36] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *ECCV*, 2016. 2

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 4

[38] Patsorn Sangkloy, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. Scribbler: Controlling deep image synthesis with sketch and color. In *CVPR*, 2017. 2

[39] 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. In *CVPR*, 2018. 2, 5

[40] Thomas Whelan, Stefan Leutenegger, R Salas-Moreno, Ben Glocker, and Andrew Davison. Elasticfusion: Dense slam without a pose graph. In *Robotics: Science and Systems*, 2015. 3

[41] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 2

[42] Jeffrey M Yau, Anitha Pasupathy, Paul J Fitzgerald, Steven S Hsiao, and Charles E Connor. Analogous intermediate shape coding in vision and touch. *Proceedings of the National Academy of Sciences*, 2009. 1

[43] Wenzhen Yuan, Siyuan Dong, and Edward H Adelson. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 2017. 2, 3

[44] Wenzhen Yuan, Rui Li, Mandayam A Srinivasan, and Edward H Adelson. Measurement of shear and slip with a gelsight tactile sensor. In *ICRA*, 2015. 2

[45] Wenzhen Yuan, Mandayam A Srinivasan, and Edward H Adelson. Estimating object hardness with a gelsight touch sensor. In *IROS*, 2016. 2

[46] Wenzhen Yuan, Shaoxiong Wang, Siyuan Dong, and Edward Adelson. Connecting look and feel: Associating the visual and tactile properties of physical materials. In *CVPR*, 2017. 2

[47] 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 *ICRA*, 2017. 2

[48] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016. 5, 6

[49] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. In *ICLR*, 2015. 7, 8

[50] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 2, 5

[51] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *NIPS*, 2017. 2