Tactile-Augmented Radiance Fields

Yiming Dou[1] Fengyu Yang[2] Yi Liu[1] Antonio Loquercio[3] Andrew Owens[1]

[1]University of Michigan [2]Yale University [3]UC Berkeley

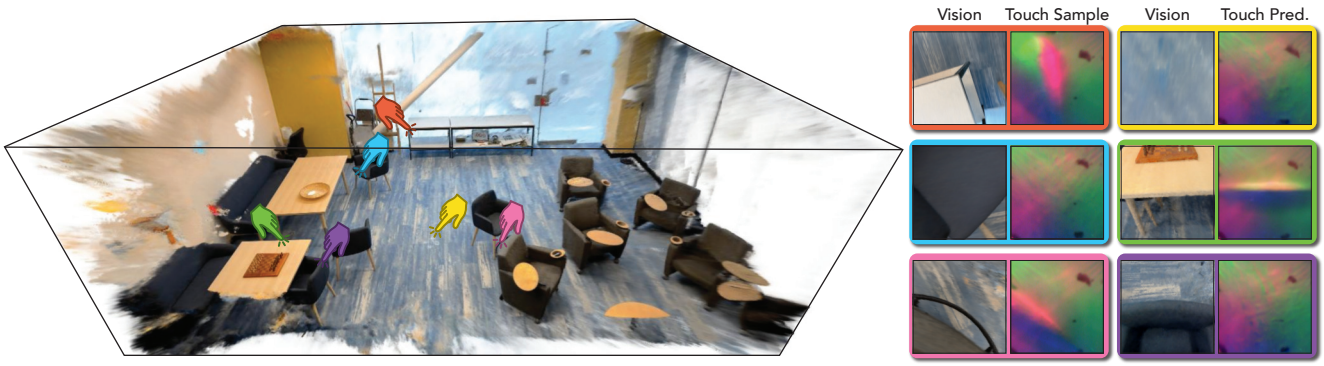

Figure 1. **Tactile-augmented radiance fields.** We capture a *tactile-augmented radiance field* (TaRF) from photos and sparsely sampled touch probes. To do this, we register the captured visual and tactile signals into a shared 3D space, then train a diffusion model to impute touch at other locations within the scene. Here, we visualize two touch probes and their (color coded) 3D positions in the scene. We also show two touch signals estimated by the diffusion model. The touch signals were collected using a vision-based touch sensor [32] that represents the touch signals as images. Please see our project page for video results.

Abstract

We present a scene representation, which we call a tactile-augmented radiance field *(TaRF), that brings vision and touch into a shared 3D space. This representation can be used to estimate the visual and tactile signals for a given 3D position within a scene. We capture a scene's TaRF from a collection of photos and sparsely sampled touch probes. Our approach makes use of two insights: (i) common vision-based touch sensors are built on ordinary cameras and thus can be registered to images using methods from multi-view geometry, and (ii) visually and structurally similar regions of a scene share the same tactile features. We use these insights to register touch signals to a captured visual scene, and to train a conditional diffusion model that, provided with an RGB-D image rendered from a neural radiance field, generates its corresponding tactile signal. To evaluate our approach, we collect a dataset of TaRFs. This dataset contains more touch samples than previous real-world datasets, and it provides spatially aligned visual signals for each captured touch signal. We demonstrate the accuracy of our cross-modal generative model and the utility of the captured visual-tactile data on several downstream tasks. Project page:* https://dou-yiming.github.io/TaRF.

1. Introduction

As humans, our ability to perceive the world relies crucially on cross-modal associations between sight and touch [19, 50]. Tactile sensing provides a detailed understanding of material properties and microgeometry, such as the intricate patterns of bumps on rough surfaces and the complex motions that soft objects make when they deform. This type of understanding, which largely eludes today's computer vision models, is a critical component of applications that require reasoning about physical contact, such as robotic locomotion [3, 24, 31, 34, 37, 38] and manipulation [6, 7, 11, 42, 60], and methods that simulate the behavior of materials [4, 13, 40, 41].

In comparison to many other modalities, collecting tactile data is an expensive and tedious process, since it requires direct physical interaction with the environment. A recent line of work has addressed this problem by having humans or robots probe the environment with touch sensors (see Table 1). Early efforts have been focused on capturing the properties of only a few objects either in simulation [16, 17, 52] or in lab-controlled settings [6, 7, 18, 28, 35, 52, 63], which may not fully convey the diversity of tactile signals in natural environments. Other works have gone beyond a

Dataset	Samples	Aligned	Scenario	Source
More Than a Feeling [7]	6.5k	✗	Tabletop	Robot
Feeling of Success [6]	9.3k	✗	Tabletop	Robot
VisGel [35]	12k	✗	Tabletop	Robot
SSVTP [28]	4.6k	✓	Tabletop	Robot
ObjectFolder 1.0 [16]	–	✓	Object	Synthetic
ObjectFolder 2.0 [17]	–	✓	Object	Synthetic
ObjectFolder Real [18]	3.7k	✗	Object	Robot
Burka et al. [5]	1.1k	✗	Sub-scene	Human
Touch and Go [56]	13.9k	✗	Sub-scene	Human
YCB-Slide* [52]	-	✓	Object	Human
Touching a NeRF [63]	1.2k	✓	Object	Robot
TaRF (Ours)	**19.3k**	✓	**Full scene**	Human

Table 1. **Dataset comparison**. We present the number of real visual-tactile pairs and whether such pairs are visually aligned, i.e., whether the visual image includes an occlusion-free view of the touched surface. *YCB-Slide has real-world touch probes but *synthetic* images rendered with CAD models of YCB objects on a white background [9].

lab setting and have collected touch from real scenes [5, 56]. However, existing datasets lack aligned visual and tactile information, since the touch sensor and the person (or robot) that holds it often occlude large portions of the visual scene (Fig. 2). These datasets also contain only a sparse set of touch signals for each scene, and it is not clear how the sampled touch signals relate to each other in 3D.

In this work, we present a simple and low-cost procedure to capture quasi-dense, scene-level, and spatially-aligned visual and touch data (Fig. 1). We call the resulting scene representation a *tactile-augmented radiance field* (TaRF). We remove the need for robotic collection by leveraging a 3D scene representation (a NeRF [39]) to synthesize a view of the surface being touched, which results in spatially aligned visual-tactile data (Fig. 2). We collect this data by mounting a touch sensor to a camera with commonly available materials (Fig. 3). To calibrate the pair of sensors, we take advantage of the fact that popular vision-based touch sensors [25, 26, 32, 48] are built on ordinary cameras. The relative pose between the vision and tactile sensors can thus be estimated using traditional methods from multi-view geometry, such as camera resectioning [20].

We use this procedure to collect a large real-world dataset of aligned visual-tactile data. With this dataset, we train a diffusion model [45, 51] to estimate touch at locations not directly probed by a sensor. In contrast to the recent work of Zhong *et al*. [63], which also estimates touch from 3D NeRF geometry, we create scene-scale reconstructions, we do not require robotic proprioception, and we use diffusion models [51]. This enables us to obtain tactile data at a much larger scale, and with considerably more diversity. Unlike previous visual-tactile diffusion work [57], we condition the model on spatially aligned visual and depth information, enhancing the generated samples' quality and their usefulness in downstream applications. After training, the diffusion model can be used to predict tactile information for novel positions in the scene. Analogous to quasi-dense stereo methods [15, 33], the diffusion model effectively propagates sparse touch samples, obtained by probing, to other visually and structurally similar regions of the scene.

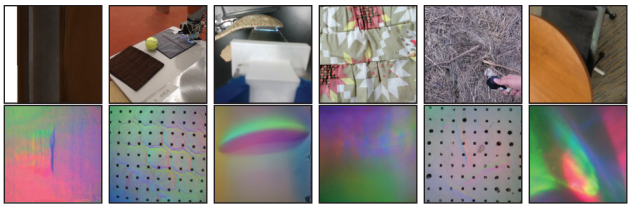

Figure 2. **Visual-tactile examples.** In contrast to the visual-tactile data captured in previous work, our approach allows us to sample unobstructed images that are spatially aligned with the touch signal, from arbitrary 3D viewpoints using a NeRF.

We evaluate our visual-tactile model's ability to accurately perform cross-modal translation using a variety of quality metrics. We also apply it to several downstream tasks, including localizing a touch within a scene and understanding material properties of the touched area. Our experiments suggest:

- Touch signals can be localized in 3D space by exploiting multi-view geometry constraints between sight and touch.
- Estimated touch measurements from novel views are not only qualitatively accurate, but also beneficial on downstream tasks.
- Cross-modal prediction models can accurately estimate touch from sight for natural scenes.
- Visually-acquired 3D scene geometry improves cross-modal prediction.

2. Related Work

Visual-tactile datasets. Previous work has either used simulators [16, 17] or robotic arms [6, 8, 18, 35, 63] for data generation. Our work is closely related to that of Zhong *et al*. [63], which uses a NeRF and captured touch data to generate a tactile field for several small objects. They use the proprioception of an expensive robot to spatially align vision and touch. In contrast, we leverage the properties of the tactile sensor and novel view synthesis to use commonly available material (a smartphone and a selfie stick) to align vision and touch. This enables the collection of a larger, scene-level, and more diverse dataset, on which we train a higher-capacity diffusion model (rather than a conditional GAN). Like several previous works [5, 56], we also collect scene-level data. In contrast to them, we spatially align the signals by registering them in a unified 3D representation, thereby increasing the prediction power of the visual-tactile generative model.

Capturing multimodal 3D scenes. Our work is related to methods that capture 3D visual reconstructions of spaces

using RGB-D data [12, 49, 55, 59] and multimodal datasets of paired 3D vision and language [1, 2, 10]. Our work is also related to recent methods that localize objects in NeRFs using joint embeddings between images and language [29] or by semantic segmentation [62]. In contrast to language supervision, touch is tied to a precise position in a scene.

3D touch sensing. A variety of works have studied the close relationship between geometry and touch, motivating our use of geometry in imputing touch. Johnson *et al*. [25, 26] proposed vision-based touch sensing, and showed that highly accurate depth can be estimated from the touch sensor using photometric stereo. Other work has estimated object-scale 3D from touch [54]. By contrast, we combine sparse estimates of touch with quasi-dense tactile signals estimated using generative models.

Cross-modal prediction of touch from sight. Recent work has trained generative models that predict touch from images. Li *et al*. [35] used a GAN to predict touch for images of a robotic arm, while Gao *et al*. [18] applied them to objects collected on a turntable. Yang *et al*. [57] used latent diffusion to predict touch from videos of humans touching objects. Our goal is different from these works: we want to predict touch signals that are spatially aligned with a visual signal, to exploit scene-specific information, and to use geometry. Thus, we use a different architecture and conditioning signal, and fit our model to examples from the same scenes at training and test time. Other work has learned joint embeddings between vision and touch [28, 36, 56, 58, 61].

3. Method

We collect visual and tactile examples from a scene and register them together with a 3D visual reconstruction to build a TaRF. Specifically, we capture a NeRF $F_\theta : (\mathbf{x}, \mathbf{r}) \mapsto (\mathbf{c}, \sigma)$ that maps a 3D point $\mathbf{x} = (x, y, z)$ and viewing direction $\mathbf{r}$ to its corresponding RGB color $\mathbf{c}$ and density σ [39]. We associate to the visual representation a *touch model* $F_\phi : \mathbf{v}_t \mapsto \boldsymbol{\tau}$ that generates the tactile signal that one would obtain by touching at the center of the image $\mathbf{v}_t$. In the following, we explain how to estimate F_θ and F_ϕ and put them into the same shared 3D space.

3.1. Capturing vision and touch signals

Obtaining a visual 3D reconstruction. We build the visual NeRF, F_θ, closely following previous work [12, 55]. A human data collector moves through a scene and records a video, covering as much of the space as possible. We then estimate camera pose using structure from motion [47] and create a NeRF using off-the-shelf packages [53]. Additional details are provided in the supplement.

Capturing and registering touch. We simultaneously collect tactile and visual signals by mounting a touch sensor on a camera (Fig. 3), obtaining synchronized touch signals $\{\boldsymbol{\tau}_i\}_{i=1}^N$ and video frames $\mathbf{v}$. We then estimate the pose of the video frames using off-the-shelf structure from motion methods [47], obtaining poses $\{p_i^v\}_{i=1}^N$. Finally, we use the calibration of the mount to obtain the poses $\{p_i^t\}_{i=1}^N$ of the tactile measurements with respect to the scene's global reference frame. As a collection device, we mount an iPhone 14 Pro to one end of a camera rod, and a DIGIT [32] touch sensor to the other end. Note that the devices can be replaced with any RGB-D camera and vision-based tactile sensor.

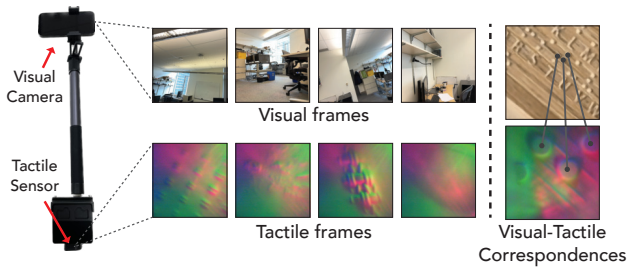

Figure 3. **Capturing setup.** (a) We record paired vision and touch signals using a camera attached to a touch sensor. (b) We estimate the relative pose between the touch sensor and the camera using correspondences between sight and touch.

Capturing setup calibration. To find the relative pose between the camera and the touch sensor (Fig. 3), we exploit the fact that arbitrary viewpoints can be synthesized from F_θ, and that ubiquitous *vision-based* touch sensors are based on perspective cameras. In these sensors, an elastomer gel is placed on the lens of a commodity camera, which is illuminated by colored lights. When the gel is pressed into an object, it deforms, and the camera records an image of the deformation; this image is used as the tactile signal. This design allows us to estimate the pose of the tactile sensor through multi-view constraints from *visual-tactile* correspondences: pixels in visual images and tactile images that are of the same physical point.

We start the calibration process by synthesizing novel views from F_θ. The views are generated at the camera location $\{p_i^v\}_{i=1}^N$, but rotated $90°$ on the x-axis. This is because the camera is approximately orthogonal to the touch sensor (see Fig. 3). Then, we manually annotate corresponding pixels between the touch measurements and the generated frames (Fig. 3). To simplify and standardize this process, we place a braille board in each scene and probe it with the touch sensor. This will generate a distinctive touch signal that is easy to localize [23].

We formulate the problem of estimating the six degrees of freedom relative pose $(\mathbf{R}, \mathbf{t})$ between the touch sensor and the generated frames as a resectioning problem [20]. We use the estimated 3D structure from the NeRF F_θ to obtain 3D points $\{\mathbf{x}_i\}_{i=1}^M$ for each of the annotated corre-

spondences. Each point has a pixel position $\mathbf{u}_i \in \mathcal{R}^2$ in the touch measurement. We find $(\mathbf{R}, \mathbf{t})$ by minimizing the reprojection error:

$$\min_{\mathbf{R},\mathbf{t}} \frac{1}{M} \sum_{i=1}^{M} \|\pi(\mathbf{K}[\mathbf{R} \mid \mathbf{t}], \mathbf{X}_i) - \mathbf{u}_i\|_1, \tag{1}$$

where π projects a 3D point using a given projection matrix, $\mathbf{K}$ are the known intrinsics of the tactile sensor's camera, and the point $\mathbf{X}_i$ is in the coordinate system of the generated vision frames. We perform the optimization on 6-15 annotated correspondences from the braille board. For robustness, we compute correspondences from multiple frames. We represent the rotation matrix using quaternions and optimize using nonlinear least-squares. Once we have $(\mathbf{R}, \mathbf{t})$ with respect to the generated frames, we can derive the relative pose between the camera and the touch sensor.

3.2. Imputing the missing touch

We use a generative model to estimate the touch signal (represented as an image from a vision-based touch sensor) for other locations within the scene. Specifically, we train a diffusion model $p_\phi(\boldsymbol{\tau} \mid \mathbf{v}, \mathbf{d}, \mathbf{b})$, where $\mathbf{v}$ and $\mathbf{d}$ are images and depth maps extracted from F_θ (see Fig. 4). We also pass as input to the diffusion model a *background* image captured by the touch sensor when it is not in contact with anything, denoted as $\mathbf{b}$. Although not essential, we have observed that this additional input empirically improves the model's performance (*e.g.*, Fig. 1 the background provides the location of defects in the gel, which appear as black dots). We train the model p_ϕ on our entire vision-touch dataset (Sec. 4).

The training of p_ϕ is divided into two stages. In the first, we pre-train a cross-modal visual-tactile encoder with self-supervised contrastive learning on our dataset. This stage, initially proposed by [23, 57], is equivalent to the self-supervised encoding pre-training that is common for image generation models [45]. We use a ResNet-50 [21] as the backbone for this contrastive model.

In the second stage, we use the contrastive model to generate the input for a conditional latent diffusion model, which is built upon Stable Diffusion [45]. A frozen pre-trained VQ-GAN [14] is used to obtain the latent representation with a spatial dimension of 64×64. We start training the diffusion model from scratch and pre-train it on the task of unconditional tactile image generation on the YCB-Slide dataset [52]. After this stage, we train the conditional generative model p_ϕ on our spatially aligned visual-tactile dataset, further fine-tuning the contrastive model end-to-end with the generation task.

At inference time, given a novel location in the 3D scene, we first render the visual signals $\hat{\mathbf{v}}$ and $\hat{\mathbf{d}}$ from NeRF, and then estimate the touch signal $\hat{\boldsymbol{\tau}}$ of the position using the diffusion model.

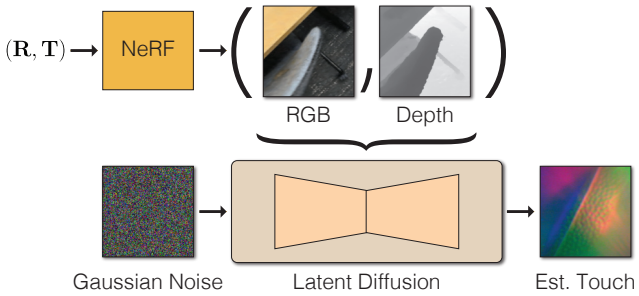

Figure 4. **Touch estimation.** We estimate the tactile signal for a given touch sensor pose $(\mathbf{R}, \mathbf{t})$. To do this, we synthesize a viewpoint from the NeRF, along with a depth map. We use conditional latent diffusion to predict the tactile signal from these inputs.

4. A 3D Visual-Tactile Dataset

In the following, we show the details of the data collection process and statistics of our dataset.

4.1. Data Collection Procedure

The data collection procedure is divided into two stages. First, we collect multiple views from the scene, capturing enough frames around the areas we plan to touch. During this stage, we collect approximately 500 frames. Next, we collect synchronized visual and touch data, maximizing the geometry and texture being touched. We then estimate the camera location of the vision frames collected in the previous two stages using off-the-shelf mapping tools [47]. After estimating the camera poses for the vision frames, the touch measurements' poses can be derived by using the mount calibration matrix. More details about the pose estimation procedure can be found in the supplement.

Finally, we associate each touch sensor with a color image by translating the sensor poses upwards by 0.4 meters and querying the NeRF with such poses. The field of view we use when querying the NeRF is $50°$. This provides us with approximately 1,500 temporally aligned vision-touch image pairs per scene. Note that this collection procedure is scalable since it does not require specific expertise or equipment and generates abundant scene-level samples.

4.2. Dataset Statistics

We collect our data in 13 ordinary scenes including two offices, a workroom, a conference room, a corridor, a tabletop, a corridor, a lounge, a room with various clothes and four outdoor scenes with interesting materials. Typically, we collect 1k to 2k tactile probes in each scene, resulting in a total of 19.3k image pairs in the dataset.

Some representative samples from the collected dataset are shown in Fig. 5. Our data includes a large variety of geometry (edges, surfaces, corners, *etc*.) and texture (plastic, clothes, snow, wood, *etc*.) of different materials in the scene. During capturing process, the collector will try to

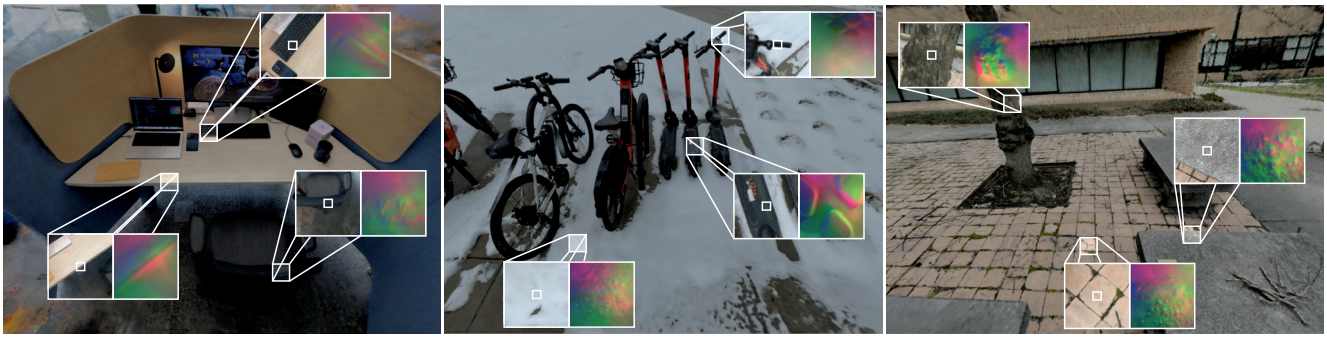

Figure 5. **Representative examples from the captured dataset**. Our dataset is obtained from nine everyday scenes, such as offices, classrooms, and kitchens. We show three such scenes in the figure above, together with samples of spatially aligned visual and tactile data. In each scene, 1k to 2k tactile probes were collected, resulting in a total of 19.3k image pairs. The data encompasses diverse geometries (edges, surfaces, corners, etc.) and textures (plastic, clothes, snow, wood, etc.) of various materials. The collector systematically probed different objects, covering areas with distinct geometry and texture using different sensor poses.

thoroughly probe various objects and cover the interesting areas with more distinguishable geometry and texture with different sensor poses. To the best of our knowledge, our dataset is the first dataset that captures full, scene-scale spatially aligned vision-touch image pairs. We provide more details about the dataset in the supplement.

5. Experiments

Leveraging the spatially aligned image and touch pairs from our dataset, we first conduct experiments on dense touch estimation. We then show the effectiveness of both the aligned data pairs and the synthesized touch signals by conducting tactile localization and material classification as two downstream tasks.

5.1. Implementation Details

NeRF. We use the Nerfacto method from Nerfstudio [53]. For each scene, we utilize approximately 2,000 images as training set, which thoroughly cover the scene from various view points. We train the network with a base learning rate of 1×10^{-2} using Adam [30] optimizer for 200,000 steps on a single NVIDIA RTX 2080 Ti GPU to achieve optimal performance.

Visual-tactile contrastive model. Following prior works [27, 57], we leverage contrastive learning methods to train a ResNet-50 [21] as visual encoder. The visual and tactile encoders share the same architecture but have different weights. We encode visual and tactile data into latent vectors in the resulting shared representation space. We set the dimension of the latent vectors to 32. Similar to CLIP [43], the model is trained on InfoNCE loss obtained from the pairwise dot products of the latent vectors. We train the model for 20 epochs by Adam [30] optimizer with a learning rate of 10^{-4} and batch size of 256 on 4 NVIDIA RTX 2080 Ti GPUs.

Visual-tactile generative model. Our implementation of the diffusion model closely follows Stable Diffusion [46], with the difference that we use a ResNet-50 to generate the visual encoding from RGB-D images for conditioning. Specifically, we also add the RGB-D images rendered from the tactile sensors' poses into the conditioning, which we refer to in Sec. 5.2 as multiscale conditioning. The model is optimized for 30 epochs by Adam [30] optimizer with a base learning rate of 10^{-5}. The learning rate is scaled by gpu number $\times$ batch size. We train the model with batch size of 48 on 4 NVIDIA A40 GPUs. At inference time, the model conducts 200 steps of denoising process with a 7.5 guidance scale. Following prior cross-modal synthesis work [44], we use reranking to improve the prediction quality. We obtain 16 samples from the diffusion model for every instance and re-rank the samples with our pretrained contrastive model. The sample with highest similarity is the final prediction.

5.2. Dense Touch Estimation

Experimental setup. We now evaluate the diffusion model's ability to generate touch images. To reduce overlap between the training and test set, we first split the frames into sequences temporally (following previous work [56]). We split them into sequences of 50 touch samples, then divide these sequences into train/validation/test with a ratio of 8/1/1. We evaluate the generated samples on Frechet Inception Distance (FID), a standard evaluation metric for cross-modal generation [56]. We also include Peak Signal to Noise Ratio (PSNR) and Structural Similarity (SSIM), though we note that these metrics are highly sensitive to spatial position of the generated content, and can be optimized by models that minimize simple pixelwise losses [22]. We also include CVTP metric proposed by prior work [57], which measures the similarity between visual and tactile embeddings of a contrastive model, analogous to

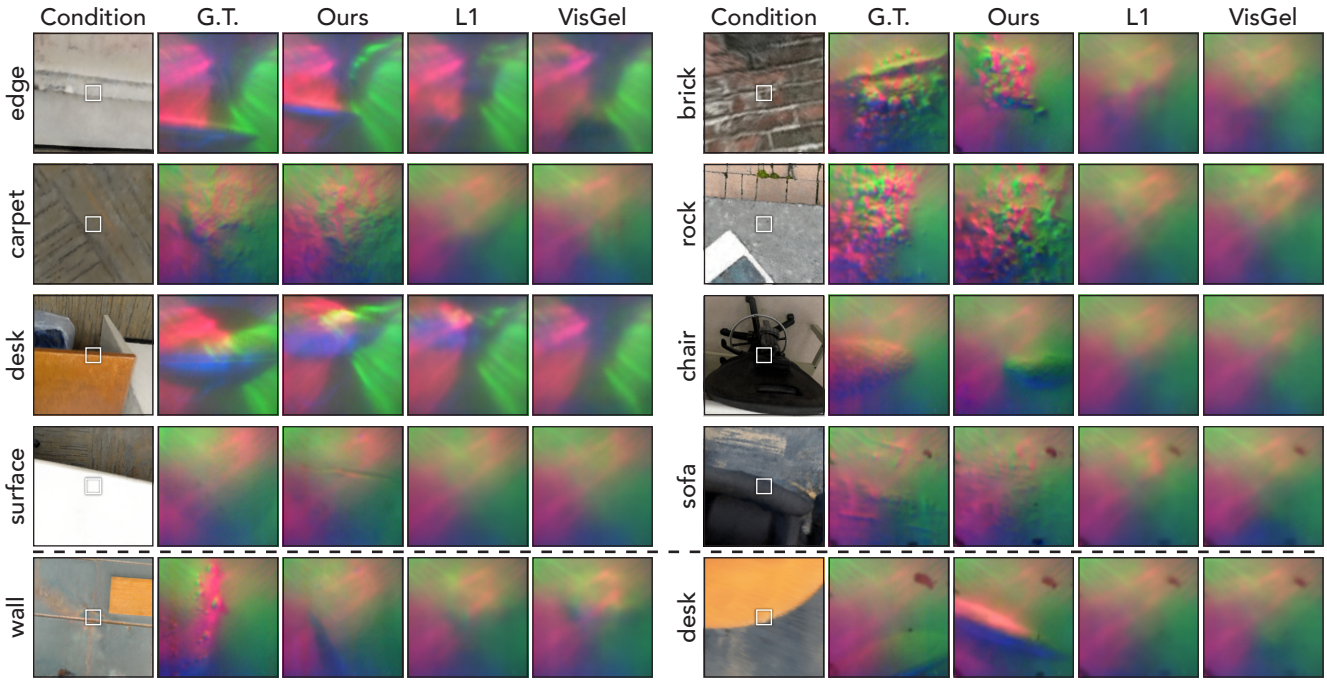

Figure 6. **Qualitative touch estimation results**. Each model is conditioned on the RGB image and depth map rendered from the NeRF (left). The white box indicates the tactile sensor's approximate field of view (which is much smaller than the full conditional image). The G.T. column shows the ground truth touch images measured from a DIGIT sensor. L1 and VisGel often generate blurry textures and inaccurate geometry. By contrast, our model better captures the features of the tactile image, *e.g.*, the rock's microgeometry and complex textures and shapes of furniture. The last row shows two failure cases of our model. In both examples, our model generates a touch image that is geometrically misaligned with the ground truth. All of the examples shown here are at least 10cm away from any training sample.

CLIP [43] score. We compare against two baselines: *VisGel*, the approach from Li et. [35], which trains a GAN for touch generation, and *L1*, a model with the same architecture of VisGel but trained to minimize an L1 loss in pixel space.

Results. As is shown in Table 2, our approach performs much better on the high-level metrics, with up to 4x lower FID and 80x higher CVTP. This indicates that our proposed diffusion model captures the distribution and characteristics of the real tactile data more effectively. On the low-level metrics (PSNR and SSIM), all methods are comparable. In particular, the L1 model slightly outperforms the other methods since the loss it is trained on is highly correlated with low-level, pixel-wise metrics. Fig. 6 qualitatively compares samples from the different models. Indeed, our generated samples exhibit enhanced details in micro-geometry of fabrics and richer textures, including snow, wood and carpeting. However, all methods fail on fine details that are barely visible in the image, such as the tree bark.

Ablation study. We evaluate the importance of the main components of our proposed touch generation approach (Table 3). Removing the conditioning on the RGB image results in the most prominent performance drop. This is expected since RGB image uniquely determines the fine-grained details of a tactile image. Removing depth image or contrastive pretraining has small effect on CVTP but results in a drop on FID. Contrastive re-ranking largely improves CVTP, indicating the necessity of obtaining multiple samples from the diffusion model. We also find that multiscale conditioning provide a small benefit on FID and CVTP.

Model	PSNR ↑	SSIM ↑	FID ↓	CVTP ↑
L1	**24.34**	**0.82**	97.05	0.01
VisGel [35]	23.66	0.81	130.22	0.03
Ours	22.84	0.72	**28.97**	**0.80**

Table 2. **Quantitative results on touch estimation for novel views.** While comparable on low-level metrics with the baselines, our approach captures the characteristics of the real tactile data more effectively, resulting in a lower FID score.

5.3. Downstream Task I: Tactile Localization

To help understand the quality of the captured TaRFs, we evaluate the performance of the contrastive model (used for conditioning our diffusion model) on the task of tactile localization. Given a tactile signal, our goal is to find the corresponding regions in a 2D image or in a 3D scene that are associated with it, *i.e.*, we ask the question: *what part of this image/scene feel like this?* We perform the following

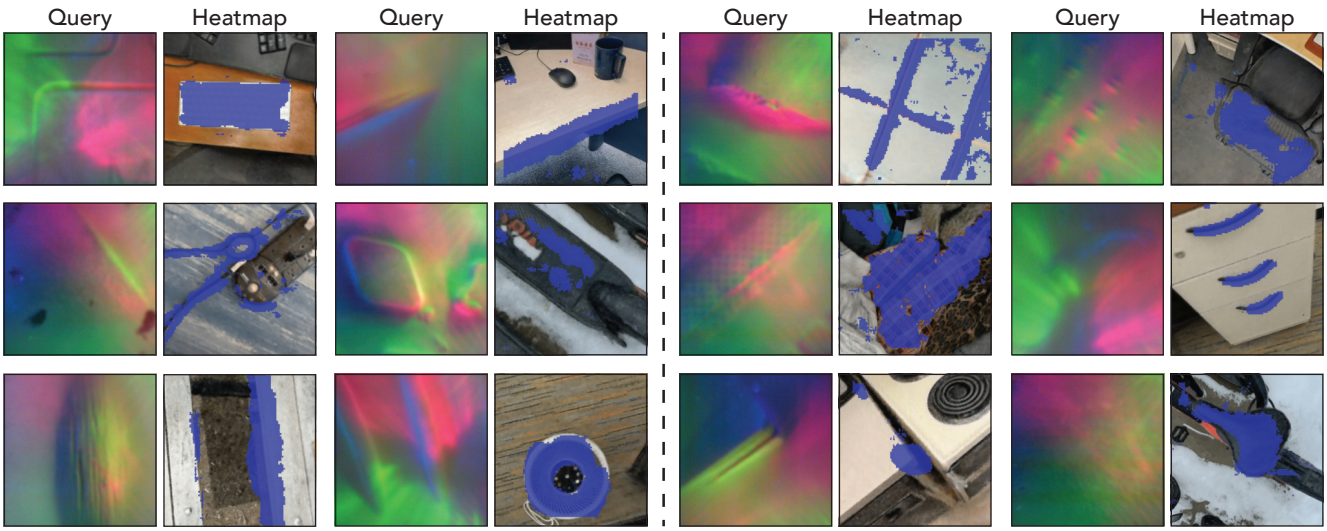

Figure 7. **Tactile localization heatmaps**. Given a tactile query image, the heatmap shows the image patches with a higher affinity to this tactile signal, as measured by a contrastive model trained on our dataset. We use a sliding window and compare each extracted patch with the touch signal. In each case, the center patch is the true position. Our model successfully captures the correlation between the two signals. This enables it to localize a variety of touch signals, including fine-grained geometry, *e.g.*, a cable or a keyboard, various types of corners and edges, and large uniform regions, such as a clothing. This ability enables our diffusion model to effectively propagate sparse touch samples to other visually and structurally similar regions of the scene.

Model variation	PSNR ↑	SSIM ↑	FID ↓	CVTP ↑
Full	22.84	**0.72**	**28.97**	**0.80**
No RGB conditioning	22.13	0.70	34.31	0.76
No depth conditioning	22.57	0.71	33.16	**0.80**
No contrastive pretraining	22.82	0.71	32.98	0.79
No re-ranking	22.92	**0.72**	29.46	0.61
No multiscale	**23.19**	**0.72**	30.89	0.77

Table 3. **Ablation study.** Since the fine-grained details of touch images can be determined from a RGB image, removing conditioning on the latter results in the largest performance drops. Re-ranking has notable impact on CVTP, indicating the necessity of obtaining multiple samples from the diffusion model.

evaluations on the test set of our dataset. Note that we run no task-specific training.

2D Localization. To determine which part of an image are associated with a given tactile measurement, we follow the same setup of SSVTP [28]. We first split the image into patches and compute their embedding. Then, we generate the tactile embedding of the input touch image. Finally, we compute the pairwise similarities between the tactile and visual embeddings, which we plot as a heatmap. As we can see in Fig. 7, our constrastive encoder can successfully capture the correlations between the visual and tactile data. For instance, the tactile embeddings of edges are associated to edges of similar shape in the visual image. Note that the majority of tactile embeddings are highly ambiguous: all edges with a similar geometry *feel* the same.

3D Localization. In 3D, the association of an image to tactile measurements becomes less ambiguous. Indeed, since tactile-visual samples are rotation-dependent, objects with similar shapes but different orientations will generate different tactile measurements. Lifting the task to 3D still does not remove all ambiguities (for example, each side of a rectangular table cannot be precisely localized). Nonetheless, we believe it to be a good fit for a quantitative evaluation since it's rare for two ambiguous parts of the scene to be touched with *exactly* the same orientation.

We use the following experimental setup for 3D localization. Given a tactile image as a query, we compute its distance in embedding space to all visual test images from the same scene. Note that all test images are associated with a 3D location. We define as ground-truth correspondences all test images at a distance of at most r from the 3D location of the test sample. We vary r to account for local ambiguities. As typical in the retrieval literature, we benchmark the performance with metric mean Average Precision (mAP).

We consider three baselines: (1) *chance*, which randomly selects corresponding samples; (2) *real*, which uses the contrastive model trained on our dataset; and (3) *real + estimated*, which trains the contrastive model on both dataset samples and a set of synthetic samples generated via the scenes' NeRF and our touch generation model. Specifically, we render a new image and corresponding touch by interpolating the position of two consecutive frames in the training dataset. This results in a training dataset for the contrastive model that is twice as large.

Dataset	$r(m)$				
	0.001	0.005	0.01	0.05	0.1
Chance	3.55	6.82	10.25	18.26	21.33
Real	12.10	22.93	32.10	50.30	57.15
Real + Est.	**14.92**	**26.69**	**36.17**	**53.62**	**60.61**

Table 4. **Quantitative results on 3D tactile localization**. We evaluate using mean Average Precision (mAP) as a metric. Training the contrastive model on our dataset of visually aligned real samples together with estimated samples from new locations in the scene results in the highest performance.

The results, presented in Table 4, demonstrate the performance benefit of employing both real and synthetic tactile pairs. Combining synthetic tactile images with the original pairs achieves highest performance on all distance thresholds. Overall, this indicates that touch measurements from novel views are not only qualitatively accurate, but also beneficial for this downstream task.

5.4. Downstream Task II: Material Classification

We investigate the efficacy of our visual-tactile dataset for understanding material properties, focusing on the task of material classification. We follow the formulation by Yang *et al*. [56], which consists of three subtasks: (i) material classification, requiring the distinction of materials among 20 possible classes; (ii) softness classification, a binary problem dividing materials as either hard or soft; and (iii) hardness classification, which requires the classification of materials as either rough or smooth.

We follow the same experimental procedure of [56]: we pretrain a contrastive model on a dataset and perform linear probing on the sub-tasks' training set. Our experiments only vary the pretraining dataset, leaving all architectural choices and hyperparameters the same. We compare against four baselines. A random classifier (*chance*); the ObjectFolder 2.0 dataset [17]; the VisGel dataset [35]; and the Touch and Go dataset [56]. Note that the touch sensor used in the test data (GelSight) differs from the one used in our dataset (DIGIT). Therefore, we use for pretraining a combination of our dataset and Touch and Go. To ensure a fair comparison, we also compare to the combination of each dataset and Touch and Go.

The findings from this evaluation, as shown in Table 5, suggest that our data improves the effectiveness of the contrastive pretraining objective, even though our data is from a different distribution. Moreover, we find that adding estimated touch probes for pretraining results in a higher performance on all the three tasks, especially the smoothness classification. This indicates that not only does our dataset covers a wide range of materials but also our diffusion model captures the distinguishable and useful patterns of different materials.

Dataset	Material	Hard/ Soft	Rough/ Smooth
Chance	18.6	66.1	56.3
ObjectFolder 2.0 [17]	36.2	72.0	69.0
VisGel [35]	39.1	69.4	70.4
Touch and Go [56]	54.7	77.3	79.4
+ ObjectFolder 2.0 [17]	54.6	87.3	84.8
+ VisGel [35]	53.1	86.7	83.6
+ Ours* (Real)	57.6	88.4	81.7
+ Ours* (Real + Estimated)	**59.0**	**88.7**	**86.1**

Table 5. **Material classification**. We show the downstream material recognition accuracy of models pre-trained on different datasets. The final rows show the performance when combining different datasets with *Touch and Go* [56]. * The task-specific training and testing datasets for this task are collected with a GelSight sensor. We note that our data comes from a different distribution, since it is collected with a DIGIT sensor [32].

6. Conclusion

In this work, we present the TaRF, a scene representation that brings vision and touch into a shared 3D space. This representation enables the generation of touch probes for novel scene locations. To build this representation, we collect the largest dataset of spatially aligned vision and touch probes.We study the utility of both the representation and the dataset in a series of qualitative and quantitative experiments and on two downstream tasks: 3D touch localization and material recognition. Overall, our work makes the first step towards giving current scene representation techniques an understanding of not only how things look, but also how they *feel*. This capability could be critical in several applications ranging from robotics to the creation of virtual worlds that look and feel like the real world.

Limitations. Since the touch sensor is based on a highly zoomed-in camera, small (centimeter-scale) errors in SfM or visual-tactile registration can lead to misalignments of several pixels between the views of the NeRF and the touch samples, which can be seen in our TaRFs. Another limitation of the proposed representation is the assumption that the scene's coarse-scale structure does not change when it is touched, an assumption that may be violated for some inelastic surfaces.

Acknowledgements. We thank Jeongsoo Park, Ayush Shrivastava, Daniel Geng, Ziyang Chen, Zihao Wei, Zixuan Pan, Chao Feng, Chris Rockwell, Gaurav Kaul and the reviewers for the valuable discussion and feedback. This work was supported by an NSF CAREER Award #2339071, a Sony Research Award, the DARPA Machine Common Sense program, and ONR MURI award N00014-21-1-2801.

References

[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, 2020. 3

[2] Peter Anderson, Ayush Shrivastava, Joanne Truong, Arjun Majumdar, Devi Parikh, Dhruv Batra, and Stefan Lee. Sim-to-real transfer for vision-and-language navigation. In *Conference on Robot Learning*, pages 671–681. PMLR, 2021. 3

[3] Jakub Bednarek, Michal Bednarek, Lorenz Wellhausen, Marco Hutter, and Krzysztof Walas. What am i touching? learning to classify terrain via haptic sensing. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7187–7193. IEEE, 2019. 1

[4] Katherine L Bouman, Bei Xiao, Peter Battaglia, and William T Freeman. Estimating the material properties of fabric from video. In *Proceedings of the IEEE international conference on computer vision*, pages 1984–1991, 2013. 1

[5] Alexander Burka. Instrumentation, data, and algorithms for visually understanding haptic surface properties. 2018. 2

[6] 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. 1, 2

[7] 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. *Robotics and Automation Letters (RA-L)*, 2018. 1, 2

[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. 2

[9] 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. IEEE, 2015. 2

[10] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European conference on computer vision*, 2020. 3

[11] Alex Church, John Lloyd, Raia Hadsell, and Nathan F Lepora. Deep reinforcement learning for tactile robotics: Learning to type on a braille keyboard. *IEEE Robotics and Automation Letters*, 5(4):6145–6152, 2020. 1

[12] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017. 3

[13] Abe Davis, Justin G Chen, and Frédo Durand. Image-space modal bases for plausible manipulation of objects in video. *ACM Transactions on Graphics (TOG)*, 2015. 1

[14] Mucong Ding, Kezhi Kong, Jingling Li, Chen Zhu, John Dickerson, Furong Huang, and Tom Goldstein. Vq-gnn: A universal framework to scale up graph neural networks using vector quantization. *Advances in Neural Information Processing Systems*, 34:6733–6746, 2021. 4

[15] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2009. 2

[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. 1, 2

[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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10598–10608, 2022. 1, 2, 8

[18] Ruohan Gao, Yiming Dou, Hao Li, Tanmay Agarwal, Jeannette Bohg, Yunzhu Li, Li Fei-Fei, and Jiajun Wu. The objectfolder benchmark: Multisensory learning with neural and real objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17276–17286, 2023. 1, 2, 3

[19] Michael SA Graziano and Charles G Gross. The representation of extrapersonal space: A possible role for bimodal, visual-tactile neurons. 1995. 1

[20] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 2, 3

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 5

[22] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5

[23] Carolina Higuera, Byron Boots, and Mustafa Mukadam. Learning to read braille: Bridging the tactile reality gap with diffusion models. *arXiv preprint arXiv:2304.01182*, 2023. 3, 4

[24] Mark A Hoepflinger, C David Remy, Marco Hutter, Luciano Spinello, and Roland Siegwart. Haptic terrain classification for legged robots. In *2010 IEEE International Conference on Robotics and Automation*, pages 2828–2833. IEEE, 2010. 1

[25] 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. 2, 3

[26] Micah K Johnson, Forrester Cole, Alvin Raj, and Edward H Adelson. Microgeometry capture using an elastomeric sensor. *ACM Transactions on Graphics (TOG)*, 2011. 2, 3

[27] Justin Kerr, Huang Huang, Albert Wilcox, Ryan Hoque, Jeffrey Ichnowski, Roberto Calandra, and Ken Goldberg. Learning self-supervised representations from vision and touch for active sliding perception of deformable surfaces. *arXiv preprint arXiv:2209.13042*, 2022. 5

[28] Justin Kerr, Huang Huang, Albert Wilcox, Ryan Hoque, Jeffrey Ichnowski, Roberto Calandra, and Ken Goldberg. Self-supervised visuo-tactile pretraining to locate and follow garment features. In *Robotics: Science and Systems*, 2023. 1, 2, 3, 7

[29] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 3

[30] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015. 5

[31] Hendrik Kolvenbach, Christian Bärtschi, Lorenz Wellhausen, Ruben Grandia, and Marco Hutter. Haptic inspection of planetary soils with legged robots. *IEEE Robotics and Automation Letters*, 4(2):1626–1632, 2019. 1

[32] Mike Lambeta, Po-Wei Chou, Stephen Tian, Brian Yang, Benjamin Maloon, Victoria Rose Most, Dave Stroud, Raymond Santos, Ahmad Byagowi, Gregg Kammerer, et al. Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robotics and Automation Letters*, 2020. 1, 2, 3, 8

[33] Maxime Lhuillier and Long Quan. A quasi-dense approach to surface reconstruction from uncalibrated images. *IEEE transactions on pattern analysis and machine intelligence*, 27(3):418–433, 2005. 2

[34] Hongyu Li, Snehal Dikhale, Soshi Iba, and Nawid Jamali. Vihope: Visuotactile in-hand object 6d pose estimation with shape completion. *IEEE Robotics and Automation Letters*, 8 (11):6963–6970, 2023. 1

[35] Yunzhu Li, Jun-Yan Zhu, Russ Tedrake, and Antonio Torralba. Connecting touch and vision via cross-modal prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10609–10618, 2019. 1, 2, 3, 6, 8

[36] 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. 3

[37] Antonio Loquercio, Ashish Kumar, and Jitendra Malik. Learning visual locomotion with cross-modal supervision. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023. 1

[38] Gabriel B Margolis, Xiang Fu, Yandong Ji, and Pulkit Agrawal. Learning physically grounded robot vision with active sensing motor policies. In *7th Annual Conference on Robot Learning*, 2023. 1

[39] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2, 3

[40] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. Visually indicated sounds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2405–2413, 2016. 1

[41] 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. 1

[42] Haozhi Qi, Brent Yi, Sudharshan Suresh, Mike Lambeta, Yi Ma, Roberto Calandra, and Jitendra Malik. General in-hand object rotation with vision and touch. *arXiv preprint arXiv:2309.09979*, 2023. 1

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 5, 6

[44] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 5

[45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 2, 4

[46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 5

[47] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3, 4

[48] Carmelo Sferrazza and Raffaello D'Andrea. Design, motivation and evaluation of a full-resolution optical tactile sensor. *Sensors*, 19(4):928, 2019. 2

[49] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV 2012*, ECCV 2012. 3

[50] Linda Smith and Michael Gasser. The development of embodied cognition: Six lessons from babies. *Artificial life*, 2005. 1

[51] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 2

[52] Sudharshan Suresh, Zilin Si, Stuart Anderson, Michael Kaess, and Mustafa Mukadam. Midastouch: Monte-carlo inference over distributions across sliding touch. In *Conference on Robot Learning*, pages 319–331. PMLR, 2023. 1, 2, 4

[53] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David Mcallister, Justin Kerr, and Angjoo Kanazawa. Nerfstudio: A modu-

lar framework for neural radiance field development. *arXiv preprint arXiv:2302.04264*, 2023. 3, 5

[54] Shaoxiong Wang, Jiajun Wu, Xingyuan Sun, Wenzhen Yuan, William T Freeman, Joshua B Tenenbaum, and Edward H Adelson. 3d shape perception from monocular vision, touch, and shape priors. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018. 3

[55] Jianxiong Xiao, Andrew Owens, and Antonio Torralba. Sun3d: A database of big spaces reconstructed using sfm and object labels. 2013. 3

[56] Fengyu Yang, Chenyang Ma, Jiacheng Zhang, Jing Zhu, Wenzhen Yuan, and Andrew Owens. Touch and go: Learning from human-collected vision and touch. *Neural Information Processing Systems (NeurIPS) - Datasets and Benchmarks Track*, 2022. 2, 3, 5, 8

[57] Fengyu Yang, Jiacheng Zhang, and Andrew Owens. Generating visual scenes from touch. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22070–22080, 2023. 2, 3, 4, 5

[58] Fengyu Yang, Chao Feng, Ziyang Chen, Hyoungseob Park, Daniel Wang, Yiming Dou, Ziyao Zeng, Xien Chen, Rit Gangopadhyay, Andrew Owens, and Alex Wong. Binding touch to everything: Learning unified multimodal tactile representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3

[59] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023. 3

[60] Zhao-Heng Yin, Binghao Huang, Yuzhe Qin, Qifeng Chen, and Xiaolong Wang. Rotating without seeing: Towards in-hand dexterity through touch. *arXiv preprint arXiv:2303.10880*, 2023. 1

[61] Wenzhen Yuan, Shaoxiong Wang, Siyuan Dong, and Edward Adelson. Connecting look and feel: Associating the visual and tactile properties of physical materials. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5580–5588, 2017. 3

[62] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15838–15847, 2021. 3

[63] Shaohong Zhong, Alessandro Albini, Oiwi Parker Jones, Perla Maiolino, and Ingmar Posner. Touching a nerf: Leveraging neural radiance fields for tactile sensory data generation. In *Conference on Robot Learning*, pages 1618–1628. PMLR, 2023. 1, 2