The OBJECTFOLDER BENCHMARK: Multisensory Learning with *Neural* and *Real* Objects

Ruohan Gao* Yiming Dou*† Hao Li* Tanmay Agarwal Jeannette Bohg Yunzhu Li
Li Fei-Fei Jiajun Wu
Stanford Univeristy

Abstract

We introduce the OBJECTFOLDER BENCHMARK*, a benchmark suite of 10 tasks for multisensory object-centric learning, centered around object recognition, reconstruction, and manipulation with sight, sound, and touch. We also introduce the* OBJECTFOLDER REAL *dataset, including the multisensory measurements for 100 real-world household objects, building upon a newly designed pipeline for collecting the 3D meshes, videos, impact sounds, and tactile readings of real-world objects. We conduct systematic benchmarking on both the 1,000 multisensory neural objects from* OBJECTFOLDER*, and the real multisensory data from* OBJECTFOLDER REAL*. Our results demonstrate the importance of multisensory perception and reveal the respective roles of vision, audio, and touch for different object-centric learning tasks. By publicly releasing our dataset and benchmark suite, we hope to catalyze and enable new research in multisensory object-centric learning in computer vision, robotics, and beyond. Project page:* `https://objectfolder.stanford.edu`

1. Introduction

Computer vision systems today excel at recognizing objects in 2D images thanks to many image datasets [3, 17, 35, 40]. There is also a growing interest in modeling an object's shape and appearance in 3D, with various benchmarks and tasks introduced [8, 28, 44, 45, 54, 61]. Despite the exciting progress, these studies primarily focus on the visual recognition of objects. At the same time, our everyday activities often involve multiple sensory modalities. Objects exist not just as *visual* entities, but they also make sounds and can be touched during interactions. The different sensory modes of an object all share the same underlying object intrinsics—its 3D shape, material property, and texture. Modeling the complete multisensory profile of objects is of great importance for many applications beyond computer vision, such as robotics, graphics, and virtual and augmented reality.

Some recent attempts have been made to combine multiple sensory modalities to complement vision for various tasks [2,6,39,58,59,63,70,73]. These tasks are often studied in tailored settings and evaluated on different datasets. As an attempt to develop assets generally applicable to diverse tasks, the OBJECTFOLDER dataset [23, 26] has been introduced and includes 1,000 neural objects with their visual, acoustic, and tactile properties. OBJECTFOLDER however has two fundamental limitations. First, no real objects are included; all multisensory data are obtained through simulation with no simulation-to-real (sim2real) calibration. Second, only a few tasks were presented to demonstrate the usefulness of the dataset and to establish the possibility of conducting sim2real transfer with the neural objects.

Consequently, we need a multisensory dataset of real objects and a robust benchmark suite for multisensory object-centric learning. To this end, we present the OBJECTFOLDER REAL dataset and the OBJECTFOLDER BENCHMARK suite, as shown in Fig. 1.

The OBJECTFOLDER REAL dataset contains multisensory data collected from 100 real-world household objects. We design a data collection pipeline for each modality: for vision, we scan the 3D meshes of objects in a dark room and record HD videos of each object rotating in a lightbox; for audio, we build a professional anechoic chamber with a tailored object platform and then collect impact sounds by striking the objects at different surface locations with an impact hammer; for touch, we equip a Franka Emika Panda robot arm with a GelSight robotic finger [18, 71] and collect tactile readings at the exact surface locations where impact sounds are collected.

The OBJECTFOLDER BENCHMARK suite consists of 10 benchmark tasks for multisensory object-centric learning, centered around object recognition, reconstruction, and manipulation. The three recognition tasks are cross-sensory retrieval, contact localization, and material classification; the three reconstruction tasks are 3D shape reconstruc-

[*] indicates equal contribution.

[†] Yiming is affiliated with Shanghai Jiao Tong University. The work was done when he was visiting Stanford University as a summer intern.

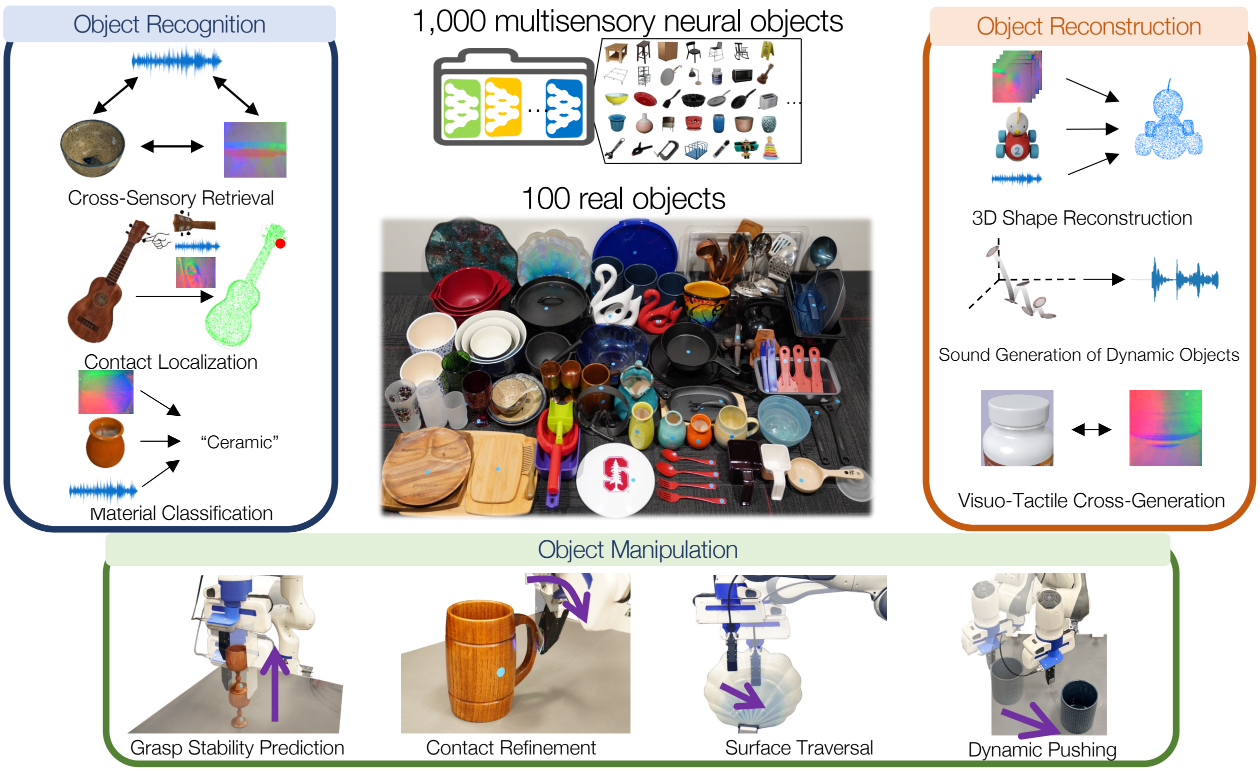


Figure 1. The OBJECTFOLDER BENCHMARK suite consists of 10 benchmark tasks for multisensory object-centric learning, centered around object recognition, reconstruction, and manipulation. Complementing the 1,000 multisensory neural objects from OBJECTFOLDER [26], we also introduce OBJECTFOLDER REAL, which contains real multisensory data collected from 100 real-world objects, including their 3D meshes, video recordings, impact sounds, and tactile readings.

tion, sound generation of dynamic objects, and visuo-tactile cross-generation; and the four manipulation tasks are grasp stability prediction, contact refinement, surface traversal, and dynamic pushing. We standardize the task setting for each task and present baseline approaches and results.

Experiments on both neural and real objects demonstrate the distinct value of sight, sound, and touch in different tasks. For recognition, vision and audio tend to be more reliable compared to touch, where the contained information is too local to recognize. For reconstruction, we observe that fusing multiple sensory modalities achieve the best results, and it is possible to hallucinate one modality from the other. This agrees with the notion of degeneracy in cognitive studies [60], which creates redundancy such that our sensory system functions even with the loss of one component. For manipulation, vision usually provides global positional information of the objects and the robot, but often suffers from occlusion. Touch, often as a good complement to vision, is especially useful to capture the accurate local geometry of the contact point.

We will open-source all code and data for OBJECTFOLDER REAL and OBJECTFOLDER BENCHMARK to facilitate research in multisensory object-centric learning.

2. Related Work

Object Datasets. A large body of work in computer vision focuses on recognizing objects in 2D images [27, 29, 30, 34]. This progress is enabled by a series of image datasets such as ImageNet [17], MS COCO [40], ObjectNet [3], and OpenImages [35]. In 3D vision, datasets like ModelNet [68] and ShapeNet [8] focus on modeling the geometry of objects but without realistic visual textures. Recently, with the popularity of neural rendering approaches [46,57], a series of 3D datasets are introduced with both realistic shape and appearance, such as CO3D [54], Google Scanned Objects [19], and ABO [14]. Unlike all datasets above that focus only on the visual modality, we also model the acoustic and tactile modalities of objects.

Our work is most related to OBJECTFOLDER [23, 26], a dataset of 1,000 neural objects with visual, acoustic, and tactile sensory data. While their multisensory data are obtained purely from simulation, we introduce the OBJECTFOLDER REAL dataset that contains real multisensory data collected from real-world household objects.

Capturing Multisensory Data from Real-World Objects. Limited prior work has attempted to capture multisensory data from the real world. Earlier work models the multisensory physical behavior of 3D objects [48] for virtual object interaction and animations. To our best knowledge, there is no large prior dataset of real object impact sounds. Datasets of real tactile data are often collected for a particular task such as robotic grasping [6,7], cross-sensory prediction [39], or from unconstrained in-the-wild settings [70]. Our OBJECTFOLDER REAL dataset is the first dataset that contains all three modalities with rich annotations to facilitate multisensory learning research with real object data.

Multisensory Object-Centric Learning. Recent work uses audio and touch in conjunction with vision for a series of new tasks, including visuo-tactile 3D reconstruction [26, 58, 59, 63], cross-sensory retrieval [2, 23], cross-modal generation [36, 39, 73], contact localization [26, 42], robotic manipulation [6, 7, 37, 38], and audio-visual learning from videos [1, 9, 11, 24, 25, 47, 74]. While they only focus on a single task of interest in tailored settings, each with a different set of objects, we present a standard benchmark suite of 10 tasks based on 1,000 neural objects from OBJECTFOLDER and 100 real objects from OBJECTFOLDER REAL for multisensory object-centric learning.

3. OBJECTFOLDER REAL

The OBJECTFOLDER dataset [26] contains 1,000 multisensory neural objects, each represented by an *Object File*, a compact neural network that encodes the object's intrinsic visual, acoustic, and tactile sensory data. Querying it with extrinsic parameters (*e.g.*, camera viewpoint and lighting conditions for vision, impact location and strength for audio, contact location and gel deformation for touch), we can obtain the corresponding sensory signal at a particular location or condition.

Though learning with these virtualized objects with simulated multisensory data is exciting, it is necessary to have a benchmark dataset of multisensory data collected from real objects to quantify the difference between simulation and reality. Having a well-calibrated dataset of real multisensory measurements allows researchers to benchmark different object-centric learning tasks on real object data without having the need to actually acquire these objects. For tasks in our benchmark suite in Sec. 4, we show results on both the neural objects from OBJECTFOLDER and the real objects from OBJECTFOLDER REAL when applicable.

Collecting real multisensory data densely from real objects is very challenging, requiring careful hardware design and tailored solutions for each sensory modality by taking into account the physical constraints (e.g., robot joint limit, kinematic constraints) in the capture system. Next, we introduce how we collect the visual (Sec. 3.1), acoustic (Sec. 3.2), and tactile (Sec. 3.3) data for the 100 real objects shown in Fig. 1. Please also visit our project page for interactive demos to visualize the captured multisensory data.

3.1. Visual Data Collection

We use an EinScan Pro HD 2020 handheld 3D Scanner[1] to scan a high-quality 3D mesh and the corresponding color texture for each object. The scanner captures highly accurate 3D features by projecting a visible light array on the object and records the texture through an attached camera. The minimum distance between two points in the scanned point cloud is $0.2\ mm$, enabling fine-grained details of the object's surface to be retained in the scanned mesh. For each object, we provide three versions of its mesh with different resolutions: 16K triangles, 64K triangles, and Full resolution (the highest number of triangles possible to achieve with the scanner). Additionally, we record an HD video of each object rotating in a lightbox with a professional camera to capture its visual appearance, as shown in Fig. 2a.

3.2. Acoustic Data Collection

We use a professional recording studio with its walls treated with acoustic melamine anechoic foam panels and the ceiling covered by absorbing acoustic ceiling tiles, as shown in Fig. 2b. The specific setup used to collect audio data varies with the object's weight and size. Most objects are placed on a circular platform made with thin strings, which minimally affects the object's vibration pattern when struck. Light objects are hung with a thin string and hit while suspended in the air. Heavy objects are placed on top of an anechoic foam panel to collect their impact sounds.

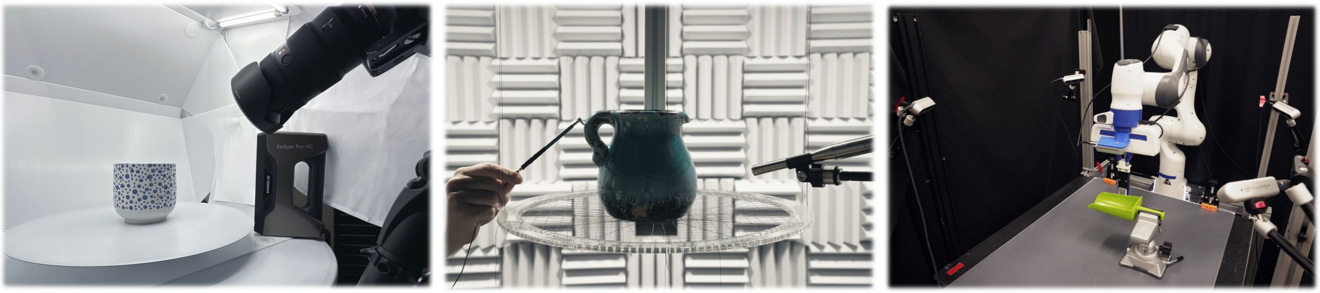
For each object, we select 30–50 points based on its scale following two criteria. First, the points should roughly cover the whole surface of the object and reveal its shape; Second, we prioritize points with specific local geometry or texture features, such as the rim/handle of a cup. For each selected point, we collect a 5-second audio clip of striking it along its normal direction with a PCB[2] impact hammer (086C01). The impact hammer is equipped with a force transducer in its tip, providing ground-truth contact forces synchronized with the audio recorded by a PCB phantom-powered free-field microphone (376A32). It is made of hardened steel, which ensures that the impacts are sharp and short enough to excite the higher-frequency modes of each object. We also record the accompanying video with a RealSense RGBD camera along with each impact sound.

3.3. Tactile Data Collection

Fig. 2c illustrates our setup for the tactile data collection. We equip a Franka Emika Panda robot arm with a GelSight touch sensor [18, 71] to automate the data collection process. GelSight sensors are vision-based tactile sensors that measure the texture and geometry of a contact surface with high spatial resolution through an elastomer and an embed-

[1] https://www.einscan.com

[2] https://www.pcb.com



(a) Visual data collection (b) Acoustic data collection (c) Tactile data collection

Figure 2. Illustration of our multisensory data collection pipeline for the ObjectFolder Real dataset. We design a tailored hardware solution for each sensory modality to collect high-fidelity visual, acoustic, and tactile data for 100 real household objects.

ded camera. We use the R1.5 GelSight tactile robot finger[3], which has a sensing area of $32 \times 24\ mm^2$.

We mount a RealSense RGBD camera at each corner of the robot frame. After camera calibration, we use the RealSense ROS package to get a point cloud estimation of the target object. We also extract a point cloud from the scanned 3D mesh of the object. In order to align the two point clouds, we first manually select four roughly corresponding points on both point clouds to provide an initial registration. Next, we use the Iterative Closest Point (ICP) [5] algorithm for point cloud alignment. We add a manual adjustment step for cases where the ICP alignment is not accurate.

We collect tactile data at the same set of surface points where the impact sounds are collected for each object. For each point of interest, we provide the robot with the target position and orientation of the GelSight robot finger; we then use position control to automatically reach the target point following the normal direction of the target point. The robot finger stops when the tactile sensor cannot deform further. We collect a video of the tactile RGB images that record the gel deformation process. We also use an in-hand camera and a third-view camera to capture two videos of the contact process for each point.

4. ObjectFolder Benchmark Suite

Our everyday activities involve the perception and manipulation of various objects. Modeling and understanding the multisensory signals of objects can potentially benefit many applications in computer vision, robotics, virtual reality, and augmented reality. The sensory streams of sight, sound, and touch all share the same underlying object intrinsics. During interactions, they often work together to reveal the object's category, 3D shape, texture, material, and physical properties.

Motivated by these observations, we introduce a suite of 10 benchmark tasks for multisensory object-centric learning, centered around *object recognition* (Sec. 4.1, 4.2, and 4.3), *object reconstruction* (Sec. 4.4, 4.5, and 4.6), and *object manipulation* (Sec. 4.7, 4.8, 4.9, and 4.10), as shown in Fig. 1. In the sections below, we first present the motivation for each task. Then, we standardize the task setting, define evaluation metrics, draw its connection to existing tasks, and develop baseline models leveraging state-of-the-art components from the literature. In the end, we show a teaser result for each task. **Please see Supp. for the complete results, baselines, and experimental setups.**

4.1. Cross-Sensory Retrieval

Motivation When seeing a wine glass, we can mentally link how it looks to how it may sound when struck or feel when touched. For machine perception, cross-sensory retrieval also plays a crucial role in understanding the relationships between different sensory modalities. While existing cross-modal retrieval benchmarks and datasets [13, 50–53] mainly focus on retrieval between images and text, we perform cross-sensory retrieval between objects' visual images, impact sounds, and tactile readings.

Task Definition. Cross-sensory retrieval requires the model to take one sensory modality as input and retrieve the corresponding data of another modality. For instance, given the sound of striking a mug, the "audio2vision" model needs to retrieve the corresponding image of the mug from a pool of images of hundreds of objects. In this benchmark, each sensory modality (vision, audio, touch) can be used as either input or output, leading to 9 sub-tasks.

Evaluation Metrics and Baselines. We measure the mean Average Precision (mAP) score, a standard metric for evaluating retrieval. We adopt several state-of-the-art methods as the baselines: 1) Canonical Correlation Analysis (CCA) [31], 2) Partial Least Squares (PLSCA) [16], 3) Deep Aligned Representations (DAR) [2], and 4) Deep Supervised Cross-Modal Retrieval (DSCMR) [75].

Teaser Results. Fig. 3 shows examples of the top retrieved instances for DAR [2], the best-performing baseline. We can see that vision and audio tend to be more reliable for retrieval, while a single touch reading usually does not contain sufficient discriminative cues to identify an object.

[3] https://www.gelsight.com

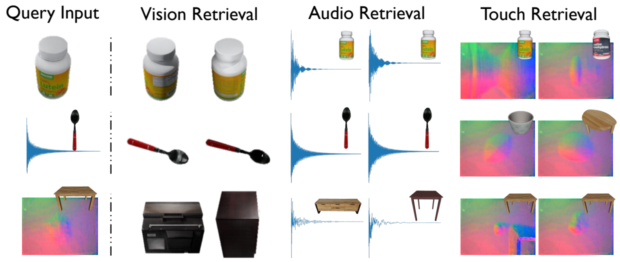


Figure 3. Examples of the top-2 retrieved instances for each modality using DAR [2], the best-performing baseline. For audio and touch retrieval, we also show an image of the object.

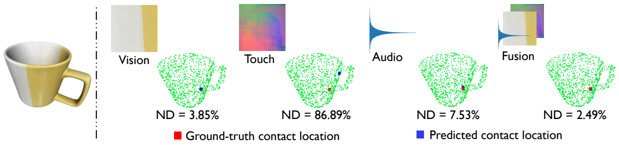


Figure 4. Contact localization results for a ceramic mug object with our multisensory contact regression model.

4.2. Contact Localization

Motivation. Localizing the contact point when interacting with an object is of great interest, especially for robot manipulation tasks. Each modality offers complementary cues: vision displays the global visual appearance of the contacting object; touch offers precise local geometry of the contact location; impact sounds at different surface locations are excited from different vibration patterns. In this benchmark task, we use or combine the object's visual, acoustic, and tactile observations for contact localization.

Task Definition. Given the object's mesh and different sensory observations of the contact position (visual images, impact sounds, or tactile readings), this task aims to predict the vertex coordinate of the surface location on the mesh where the contact happens.

Evaluation Metrics and Baselines. We use the average Normalized Distance (ND) as our metric, which measures the distance between the predicted contact position and the ground-truth position normalized by the largest distance of two points on the object's surface. We evaluate an existing baseline Point Filtering [26, 41], where the contact position is recursively filtered out based on both the multisensory observations and the relative pose between consecutive contacts. This method performs very well but heavily relies on knowing the relative pose of the series of contacts, which might be a strong assumption in practice. Therefore, we also propose a new differentiable end-to-end learning baseline for contact localization—Multisensory Contact Regression (MCR), which takes the object mesh and multisensory observations as input to regress the contact position directly.

Teaser Results. Fig. 4 shows an example result for a ceramic mug object with our MCR baseline. While vision and audio perform similarly, a single touch cannot easily locate where the contact is. Combining the three sensory modalities leads to the best result.

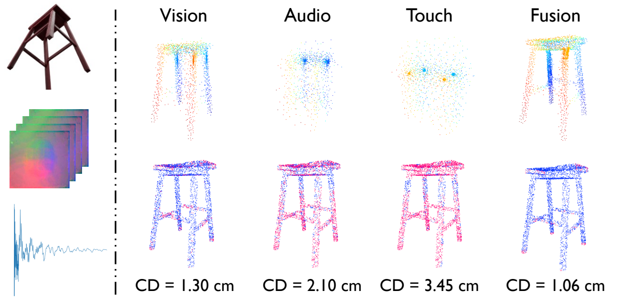


Figure 5. 3D reconstruction results of a wooden chair object. The top/bottom row shows the point cloud reconstructions and the error over ground-truth points, respectively. Red indicates poorly-reconstructed areas; CD denotes Chamfer Distance.

4.3. Material Classification

Motivation. Material is an intrinsic property of an object, which can be perceived from different sensory modalities. For example, a ceramic object usually looks glossy, sounds crisp, and feels smooth. In this task, we predict an object's material category based on its multisensory observations.

Task Definition. All objects are labeled by seven material types: ceramic, glass, wood, plastic, iron, polycarbonate, and steel. The task is formulated as a single-label classification problem. Given an RGB image, an impact sound, a tactile image, or their combination, the model must predict the correct material label for the target object.

Evaluation Metrics and Baselines. We report the classification accuracy and use two baselines: 1) ResNet [30] and 2) FENet [69], which uses a different base architecture.

Teaser Results. We conduct material classification on both neural and real objects. Fusing different modalities largely improves the material classification accuracy. We also finetune the model trained on neural objects with only a few real-world measurements and achieve 6% accuracy gain in classifying real objects.

4.4. 3D Shape Reconstruction

Motivation. While single-image shape reconstruction has been widely studied [12, 44, 49, 73], humans don't use vision alone to perceive the shape of objects. For example, we can touch an object's surface to sense its local details, or even knock and listen to the sound it makes to estimate its scale. The effective fusion of complementary multisensory information plays a vital role in 3D shape reconstruction, which we study in this benchmark task.

Task Definition. Given an RGB image of an object, a sequence of tactile readings from the object's surface, or a

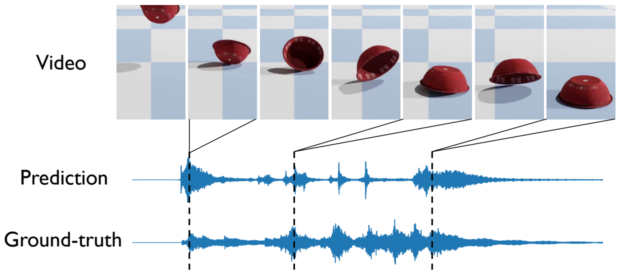


Figure 6. Example results of sound generation for a falling steel bowl object with the RegNet [10] baseline.

sequence of impact sounds of striking its surface locations, the task is to reconstruct the point cloud of the target object given combinations of these multisensory observations. This task is related to prior efforts on visuo-tactile 3D reconstruction [55, 58, 59, 62], but here we use all three sensory modalities and study their respective roles.

Evaluation Metrics and Baselines. We report Chamfer Distance [4] between the reconstructed and the ground-truth point cloud, a widely used metric to evaluate the quality of shape reconstruction. We use two state-of-the-art methods as our baseline models: 1) Mesh Deformation Network (MDN) [59], which is based on deforming the vertices of an initial mesh through a graph convolutional neural network, and 2) Point Completion Network (PCN) [26, 72], which predicts the whole point cloud from latent features or incomplete point cloud constructed from local observations.

Teaser Results. For 3D reconstruction, our observation is that vision usually provides global yet coarse information, audio indicates the object's scale, and touch provides precise local geometry of the object's surface. Fig. 5 shows an example of a wooden chair object. Both qualitative and quantitative results show that the three modalities make up for each other's deficiencies, and achieve the best reconstruction results when fused together.

4.5. Sound Generation of Dynamic Objects

Motivation Objects make unique sounds during interactions. When an object falls, we can anticipate how it sounds by inferring from its visual appearance and movement. In this task, we aim to generate the sound of dynamic objects based on videos displaying their moving trajectories.

Task Definition. Given a video clip of a falling object, the goal of this task is to generate the corresponding sound based on the visual appearance and motion of the object. The generated sound must match the object's intrinsic properties (e.g., material type) and temporally align with the object's movement in the given video. This task is related to prior work on sound generation from in-the-wild videos [10, 32, 76], but here we focus more on predicting soundtracks that closely match the object dynamics.

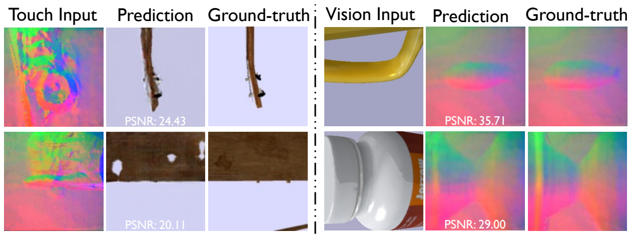


Figure 7. Examples of Touch2Vision (left) and Vision2Touch (right) cross-generation results with the VisGel [39] baseline.

Evaluation Metrics and Baselines. We use the following metrics for evaluating the sound generation quality: 1) STFT-Distance, which measures the Euclidean distance between the ground truth and predicted spectrograms, 2) Envelope Distance, which measures the Euclidean distance between the envelopes of the ground truth and the predicted signals, and 3) CDPAM [43], which measures the perceptual audio similarity. We use two state-of-the-art methods as our baselines: RegNet [10] and SpecVQGAN [32].

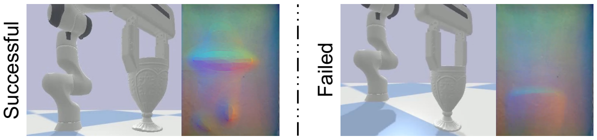
Teaser Results. Fig. 6 shows an example of the predicted sound for a falling plate. We observe that the generated sound matches well with the ground-truth sound of the object perceptually, but it is challenging to predict the exact alignment that matches the object's motion.

4.6. Visuo-Tactile Cross-Generation

Motivation. When we touch an object that is visually occluded (e.g., searching for a wallet from a backpack), we can often anticipate its visual textures and geometry merely based on the feeling on our fingertips. Similarly, we may imagine the feeling of touching an object purely from a glimpse of its visual appearance and vice-versa. To realize this intuition, we study the visuo-tactile cross-generation task initially proposed in [39].

Task Definition. We can either predict touch from vision or vision from touch, leading to two subtasks: 1) Vision2Touch: Given an image of a local region on the object's surface, predict the corresponding tactile RGB image that aligns with the visual image patch in both position and orientation; and 2) Touch2Vision: Given a tactile reading on the object's surface, predict the corresponding local image patch where the contact happens.

Evaluation Metrics and Baselines. Both the visual and tactile sensory data are represented by RGB images. Therefore, we evaluate the prediction performance for both subtasks using Peak Signal to Noise Ratio (PSNR) and Structural Similarity (SSIM) — widely used metrics for assessing image prediction quality. We use two image-to-image translation methods as our baselines: 1) Pix2Pix [33], which is a general-purpose conditional GAN framework, and 2)



CHANCE	V	T	V + T
50%	87.4%	93.4%	99.4%

Figure 8. Grasp stability prediction results with a wine glass. We show an example of a successful grasp (left) and one of a failed grasp (right). The table shows the prediction accuracy with V and T denoting using vision and/or touch, respectively.

VisGel [39], which is a variant of Pix2Pix that is specifically designed for cross-sensory prediction.

Teaser Results. Fig. 7 shows some examples of visuo-tactile cross-generation. Very accurate touch signals can be reconstructed from local views of the objects, while visual image patches generated from tactile input tend to lose surface details. We suspect this is because different objects often share similar local patterns, making it ambiguous to invert visual appearance from a single tactile reading.

4.7. Grasp-Stability Prediction

Motivation. Grasping an object is inherently a multisensory experience. When we grasp an object, vision helps us quickly localize the object, and touch provides an accurate perception of the local contact geometry. Both visual and tactile senses are useful for predicting the stability of robotic grasping, which has been studied in prior work with various task setups [7, 56, 66].

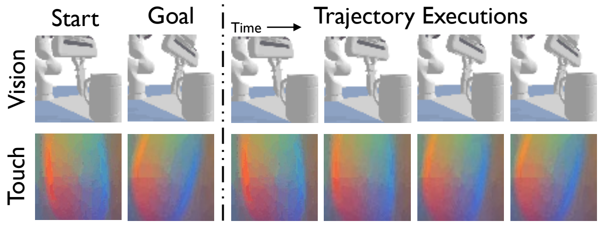
Task Definition. The goal is to predict whether a robotic gripper can successfully grasp and stably hold an object between its left and right fingers based on either an image of the grasping moment from an externally mounted camera, a tactile RGB image obtained from the GelSight robot finger, or their combination. The grasp is considered failed if the grasped object slips by more than 3 cm.

Evaluation Metrics and Baselines. We report the accuracy of grasp stability prediction. We implement TACTO [66] as the baseline method, which uses a ResNet-18 [30] network for feature extraction from the visual and tactile RGB images to predict the grasp stability.

Teaser Results. We show a successful and a failed grasp for a wine glass in Fig. 8. Vision and touch are both helpful in predicting grasp stability, and combining the two sensory modalities leads to the best result.

4.8. Contact Refinement

Motivation. When seeing a cup, we can instantly analyze its shape and structure, and decide to put our fingers around its handle to lift it. We often slightly adjust the orientations of our fingers to achieve the most stable pose for grasping. For robots, locally refining how it contacts an object is of great practical importance. We define this new task as *contact refinement*, which can potentially be a building block for many dexterous manipulation tasks.



	V	T	V + T
SR ↑	0.86	0.83	0.88
AE ↓ (°)	0.38	0.56	0.34

Figure 9. Contact refinement results of a wooden cup object. From left to right, we show the start and goal observations for both vision (top) and touch (bottom), and the actual trajectory executions. The table shows the success rate (SR) and the angle error (AE) for using vision (V), touch (T), or its combination.

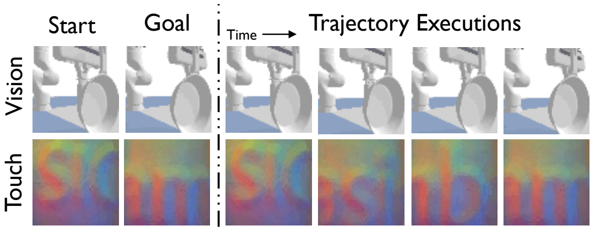
Task Definition. Given an initial pose of the robot finger, the task is to change the finger's orientation to contact the point with a different target orientation. Each episode is defined by the following: the contact point, the start orientation of the robot finger along the vertex normal direction of the contact point, and observations from the target finger orientation in the form of either a third view camera image, a tactile RGB image, or both. We use a continuous action space over the finger rotation dimension. The task is successful if the finger reaches the target orientation within 15 action steps with a tolerance of 1°.

Evaluation Metrics and Baselines. We evaluate using the following metrics: 1) success rate (SR), which is the fraction of successful trials, and 2) average Angle Error (AE) across all test trials. Model Predictive Control (MPC) [20, 22, 64] has been shown to be a powerful framework for planning robot actions. Therefore, we implement Multisensory-MPC as our baseline, which uses SVG [65] for future frame prediction, and Model Predictive Path Integral Control (MPPI) [67] for training the control policy.

Teaser Results. Fig. 9 shows a trajectory execution example for using both vision and touch. We can obtain an 88% success rate and average angle error of 0.17° by combining both modalities using our Multisensory-MPC baseline.

4.9. Surface Traversal

Motivation. When a robot's finger first contacts a position on an object, it may not be the desired surface location. Therefore, efficiently traversing from the first contact



	V	T	V + T
SR ↑	0.26	0.54	0.80
PE ↓ (mm)	1.77	3.32	0.84

Figure 10. Trajectory executions examples for surface traversal with an iron pan. The table shows the success rate (SR) and average position error (PE) for using vision (V) and/or touch (T).

point to the target location is a prerequisite for performing follow-up actions or tasks. We name this new task *surface traversal*, where we combine visual and tactile sensing to efficiently traverse to the specified target location given a visual and/or tactile observation of the starting location.

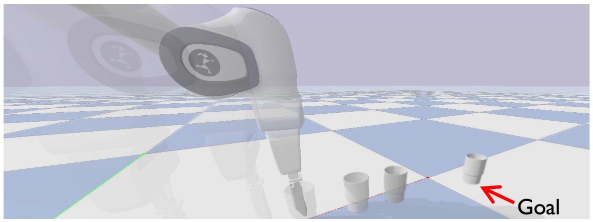
Task Definition. Given an initial contacting point, the goal of this task is to plan a sequence of actions to move the robot finger horizontally or vertically in the contact plane to reach another target location on the object's surface. Each episode is defined by the following: the initial contact point, and observations of the target point in the form of either a third-view camera image, a tactile RGB image, or both. The task is successful if the robot finger reaches the target point within 15 action steps with a tolerance of 1 mm.

Evaluation Metrics and Baselines. We report the following two metrics: 1) success rate (SR), and 2) average position error (PE), which is the average distance between the final location of the robot finger on the object's surface and the target location. We use the same Multisensory-MPC baseline as in the contact refinement task.

Teaser Results. Fig. 10 shows the surface traversal results with an iron pan, where the back of the pan has a sequence of letters. The Multisensory-MPC model can successfully traverse from the start location to the goal location. We observe significant gains when combining vision and touch, achieving a success rate of 80%.

4.10. Dynamic Pushing

Motivation. To push an object to a target location, we use vision to gauge the distance and tactile feedback to control the force and orientation. For example, in curling, the player sees and decides on the stone's target, holds its handle to push, and lightly turns the stone in one direction or the other upon release. Both visual and tactile signals play a crucial role in a successful delivery. We name this task *dynamic pushing*, which is related to prior work on dynamic adaptation for pushing [21] with only vision.



	V	T	V + T
PE ↓ (cm)	23.81	21.76	17.63

Figure 11. Examples of dynamic pushing. The table shows the average position error (PE) for using vision (V) and/or touch (T) with a rinsing cup.

Task Definition. Given example trajectories of pushing different objects together with their corresponding visual and tactile observations, the goal of this task is to learn a forward dynamics model that can quickly adapt to novel objects with a few contextual examples. With the learned dynamics model, the robot is then tasked to push the objects to new goal locations.

Evaluation Metrics and Baselines. We report the average position error (PE) across all test trials. For the baseline, we use a ResNet-18 network for feature extraction and a self-attention mechanism for modality fusion to learn the forward dynamics model. We use a sampling-based optimization algorithm (i.e., cross-entropy method [15]) to obtain the control signal.

Teaser Results. Fig. 11 shows an example of pushing a novel test object to a new goal location. Vision and touch are both useful for learning object dynamics, and combining the two sensory modalities leads to the best results.

5. Conclusion

We presented the OBJECTFOLDER BENCHMARK, a suite of 10 benchmark tasks centered around object recognition, reconstruction, and manipulation to advance research on multisensory object-centric learning. We also introduced OBJECTFOLDER REAL, the first dataset that contains all visual, acoustic, and tactile real-world measurements of 100 real household objects. We hope our new dataset and benchmark suite can serve as a solid building block to enable further research and innovations in multisensory object modeling and understanding.

Acknowledgments. We thank Samuel Clarke, Miaoya Zhong, Mark Rau, Hong-Xing Yu, and Samir Agarwala for helping with the data collection, and thank Doug James, Zilin Si, Fengyu Yang, and Yen-Yu Chang for helpful discussions. This work is in part supported by Amazon, NSF CCRI #2120095, NSF RI #2211258, ONR MURI N00014-22-1-2740, AFOSR YIP FA9550-23-1-0127, the Stanford Institute for Human-Centered AI (HAI), the Toyota Research Institute (TRI), Adobe, and Meta.

References

[1] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *ICCV*, 2017. 3

[2] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. See, hear, and read: Deep aligned representations. *arXiv preprint arXiv:1706.00932*, 2017. 1, 3, 4, 5

[3] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *NeurIPS*, 2019. 1, 2

[4] Harry G Barrow, Jay M Tenenbaum, Robert C Bolles, and Helen C Wolf. Parametric correspondence and Chamfer matching: Two new techniques for image matching. In *Proceedings: Image Understanding Workshop*, pages 21–27, 1977. 6

[5] Paul J Besl and Neil D McKay. Method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14:239–256, 1992. 4

[6] Roberto Calandra, Andrew Owens, Dinesh Jayaraman, Justin Lin, Wenzhen Yuan, Jitendra Malik, Edward H Adelson, and Sergey Levine. More than a feeling: Learning to grasp and regrasp using vision and touch. *RA-L*, 2018. 1, 3

[7] 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 *CoRL*, 2017. 3, 7

[8] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. ShapeNet: An information-rich 3D model repository. *arXiv preprint arXiv:1512.03012*, 2015. 1, 2

[9] Changan Chen, Ruohan Gao, Paul Calamia, and Kristen Grauman. Visual acoustic matching. In *CVPR*, 2022. 3

[10] Peihao Chen, Yang Zhang, Mingkui Tan, Hongdong Xiao, Deng Huang, and Chuang Gan. Generating visually aligned sound from videos. *IEEE Transactions on Image Processing*, 2020. 6

[11] Ziyang Chen, David F Fouhey, and Andrew Owens. Sound localization by self-supervised time delay estimation. In *ECCV*, 2022. 3

[12] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3D-R2N2: A unified approach for single and multi-view 3D object reconstruction. In *ECCV*, 2016. 5

[13] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. NUS-WIDE: A real-world web image database from national university of singapore. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, 2009. 4

[14] Jasmine Collins, Shubham Goel, Achleshwar Luthra, Leon Xu, Kenan Deng, Xi Zhang, Tomas F Yago Vicente, Himanshu Arora, Thomas Dideriksen, Matthieu Guillaumin, and Jitendra Malik. ABO: Dataset and benchmarks for real-world 3D object understanding. *arXiv preprint arXiv:2110.06199*, 2021. 2

[15] Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. A tutorial on the cross-entropy method. *Annals of Operations Research*, 2005. 8

[16] Sijmen de Jong, Barry M. Wise, and N. Lawrence Ricker. Canonical partial least squares and continuum power regression. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 2001. 4

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 1, 2

[18] Siyuan Dong, Wenzhen Yuan, and Edward H Adelson. Improved gelsight tactile sensor for measuring geometry and slip. In *IROS*, 2017. 1, 3

[19] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3D scanned household items. In *ICRA*, 2022. 2

[20] Frederik Ebert, Chelsea Finn, Sudeep Dasari, Annie Xie, Alex Lee, and Sergey Levine. Visual foresight: Model-based deep reinforcement learning for vision-based robotic control. *arXiv preprint arXiv:1812.00568*, 2018. 7

[21] Ben Evans, Abitha Thankaraj, and Lerrel Pinto. Context is everything: Implicit identification for dynamics adaptation. In *ICRA*, 2022. 8

[22] Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. In *ICRA*, 2017. 7

[23] 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, 3

[24] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *ECCV*, 2018. 3

[25] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *ICCV*, 2019. 3

[26] Ruohan Gao, Zilin Si, Yen-Yu Chang, Samuel Clarke, Jeannette Bohg, Li Fei-Fei, Wenzhen Yuan, and Jiajun Wu. ObjectFolder 2.0: A multisensory object dataset for sim2real transfer. In *CVPR*, 2022. 1, 2, 3, 5, 6

[27] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 2

[28] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *CVPR*, 2022. 1

[29] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 2

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2, 5, 7

[31] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936. 4

[32] Vladimir Iashin and Esa Rahtu. Taming visually guided sound generation. In *BMVC*, 2021. 6

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

[34] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 2017. 2

[35] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *IJCV*, 2020. 1, 2

[36] Jet-Tsyn Lee, Danushka Bollegala, and Shan Luo. "Touching to see" and "seeing to feel": Robotic cross-modal sensory data generation for visual-tactile perception. In *ICRA*, 2019. 3

[37] 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. 3

[38] Hao Li, Yizhi Zhang, Junzhe Zhu, Shaoxiong Wang, A. Michelle Lee, Huazhe Xu, Edward Adelson, Fei-Fei Li, Ruohan Gao, and Jiajun Wu. See, hear, and feel: Smart sensory fusion for robotic manipulation. In *CoRL*, 2022. 3

[39] Yunzhu Li, Jun-Yan Zhu, Russ Tedrake, and Antonio Torralba. Connecting touch and vision via cross-modal prediction. In *CVPR*, 2019. 1, 3, 6, 7

[40] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 1, 2

[41] Jun S Liu and Rong Chen. Sequential monte carlo methods for dynamic systems. *Journal of the American Statistical Association*, 93(443):1032–1044, 1998. 5

[42] Shan Luo, Wenxuan Mou, Kaspar Althoefer, and Hongbin Liu. Localizing the object contact through matching tactile features with visual map. In *ICRA*, 2015. 3

[43] Pranay Manocha, Zeyu Jin, Richard Zhang, and Adam Finkelstein. CDPAM: Contrastive learning for perceptual audio similarity. In *ICASSP*, 2021. 6

[44] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, 2019. 1, 5

[45] 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. In *ECCV*, 2020. 1

[46] Thomas Neff, Pascal Stadlbauer, Mathias Parger, Andreas Kurz, Joerg H Mueller, Chakravarty R Alla Chaitanya, Anton Kaplanyan, and Markus Steinberger. DONeRF: Towards real-time rendering of compact neural radiance fields using depth oracle networks. In *EGSR*, 2021. 2

[47] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*, 2018. 3

[48] Dinesh K Pai, Kees van den Doel, Doug L James, Jochen Lang, John E Lloyd, Joshua L Richmond, and Som H Yau. Scanning physical interaction behavior of 3D objects. In *SIGGRAPH*, 2001. 3

[49] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *CVPR*, 2019. 5

[50] Yuxin Peng, Xin Huang, and Yunzhen Zhao. An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges. *TCSVT*, 2017. 4

[51] Yuxin Peng, Jinwei Qi, and Yuxin Yuan. Modality-specific cross-modal similarity measurement with recurrent attention network. *TIP*, 2018. 4

[52] Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Nikhil Rasiwasia, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. On the role of correlation and abstraction in cross-modal multimedia retrieval. *IEEE TPAMI*, 36(03):521–535, 2014. 4

[53] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT Workshops*, 2010. 4

[54] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3D: Large-scale learning and evaluation of real-life 3D category reconstruction. In *ICCV*, 2021. 1, 2

[55] Lukas Rustler, Jens Lundell, Jan Kristof Behrens, Ville Kyrki, and Matej Hoffmann. Active visuo-haptic object shape completion. *RA-L*, 2022. 6

[56] Zilin Si, Zirui Zhu, Arpit Agarwal, Stuart Anderson, and Wenzhen Yuan. Grasp stability prediction with sim-to-real transfer from tactile sensing. In *IROS*, 2022. 7

[57] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3D-structure-aware neural scene representations. In *NeurIPS*, 2019. 2

[58] Edward Smith, David Meger, Luis Pineda, Roberto Calandra, Jitendra Malik, Adriana Romero Soriano, and Michal Drozdzal. Active 3D shape reconstruction from vision and touch. *NeurIPS*, 2021. 1, 3, 6

[59] Edward J Smith, Roberto Calandra, Adriana Romero, Georgia Gkioxari, David Meger, Jitendra Malik, and Michal Drozdzal. 3D shape reconstruction from vision and touch. In *NeurIPS*, 2020. 1, 3, 6

[60] Linda Smith and Michael Gasser. The development of embodied cognition: Six lessons from babies. *Artificial Life*, 2005. 2

[61] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. Pix3D: Dataset and methods for single-image 3D shape modeling. In *CVPR*, 2018. 1

[62] Sudharshan Suresh, Zilin Si, Joshua G Mangelson, Wenzhen Yuan, and Michael Kaess. Efficient shape mapping through dense touch and vision. *arXiv preprint arXiv:2109.09884*, 2021. 6

[63] Sudharshan Suresh, Zilin Si, Joshua G Mangelson, Wenzhen Yuan, and Michael Kaess. ShapeMap 3-D: Efficient shape mapping through dense touch and vision. In *ICRA*, 2022. 1, 3

[64] Stephen Tian, Frederik Ebert, Dinesh Jayaraman, Mayur Mudigonda, Chelsea Finn, Roberto Calandra, and Sergey Levine. Manipulation by feel: Touch-based control with deep predictive models. In *ICRA*, 2019. 7

[65] Ruben Villegas, Arkanath Pathak, Harini Kannan, Dumitru Erhan, Quoc V Le, and Honglak Lee. High fidelity video prediction with large stochastic recurrent neural networks. *NeurIPS*, 2019. 7

[66] Shaoxiong Wang, Mike Lambeta, Po-Wei Chou, and Roberto Calandra. TACTO: A fast, flexible and open-source simulator for high-resolution vision-based tactile sensors. *arXiv preprint arXiv:2012.08456*, 2020. 7

[67] Grady Williams, Paul Drews, Brian Goldfain, James M. Rehg, and Evangelos A. Theodorou. Aggressive driving with model predictive path integral control. In *ICRA*, 2016. 7

[68] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3D ShapeNets: A deep representation for volumetric shapes. In *CVPR*, 2015. 2

[69] Yong Xu, Feng Li, Zhile Chen, Jinxiu Liang, and Yuhui Quan. Encoding spatial distribution of convolutional features for texture representation. In *NeurIPS*, 2021. 5

[70] Fengyu Yang, Chenyang Ma, Jiacheng Zhang, Jing Zhu, Wenzhen Yuan, and Andrew Owens. Touch and go: Learning from human-collected vision and touch. In *NeurIPS Datasets and Benchmarks Track*, 2022. 1, 3

[71] Wenzhen Yuan, Siyuan Dong, and Edward H Adelson. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 2017. 1, 3

[72] Wentao Yuan, Tejas Khot, David Held, Christoph Mertz, and Martial Hebert. PCN: Point completion network. In *3DV*, 2018. 6

[73] Zhoutong Zhang, Qiujia Li, Zhengjia Huang, Jiajun Wu, Joshua B Tenenbaum, and William T Freeman. Shape and material from sound. In *NeurIPS*, 2017. 1, 3, 5

[74] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *ECCV*, 2018. 3

[75] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. Deep supervised cross-modal retrieval. In *CVPR*, 2019. 4

[76] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L Berg. Visual to sound: Generating natural sound for videos in the wild. In *CVPR*, 2018. 6