MIT Open Access Articles

Dynamic Modeling of Hand-Object Interactions via Tactile Sensing

Citation: Zhang, Qiang, Li, Yunzhu, Luo, Yiyue, Shou, Wan, Foshey, Michael et al. 2021. "Dynamic Modeling of Hand-Object Interactions via Tactile Sensing." 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).

As Published: 10.1109/IROS51168.2021.9636361

Publisher: Institute of Electrical and Electronics Engineers (IEEE)

Persistent URL: https://hdl.handle.net/1721.1/143794

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Massachusetts Institute of Technology

Dynamic Modeling of Hand-Object Interactions via Tactile Sensing

Qiang Zhang*, Yunzhu Li*, Yiyue Luo, Wan Shou, Michael Foshey,
Junchi Yan, Joshua B. Tenenbaum, Wojciech Matusik, and Antonio Torralba
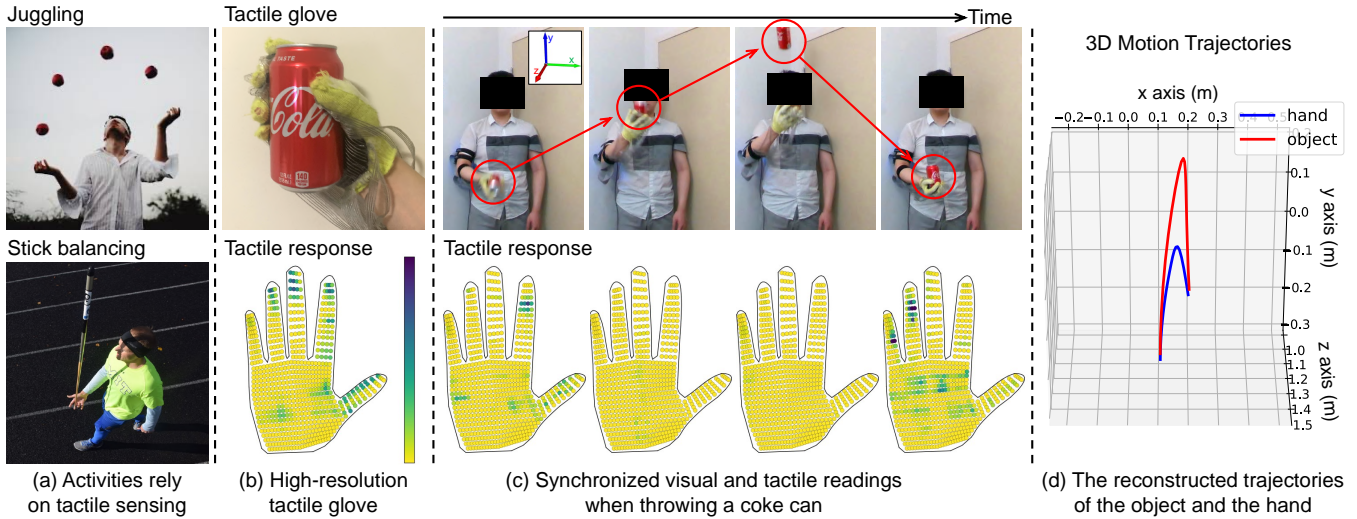
Fig. 1: **Hand-Object Interactions Using a Tactile Glove.** (a) Daily activities involving hand-object interactions require tactile sensing and an intuitive understanding of real-world dynamics. (b) We employ a high-resolution tactile glove [1] to obtain the tactile signal when interacting with objects. (c) When throwing an object in the air and catching it, different tactile patterns appear during the process. (d) We model the dynamics of this interaction and reconstruct the 3D trajectories of both the object and the hand using only tactile data.

Abstract— Tactile sensing is critical for humans to perform everyday tasks. While significant progress has been made in analyzing object grasping from vision, it remains unclear how we can utilize tactile sensing to reason about and model the dynamics of hand-object interactions. In this work, we employ a high-resolution tactile glove to perform four different interactive activities on a diversified set of objects. We build our model on a cross-modal learning framework and generate the labels using a visual processing pipeline to supervise the tactile model, which can then be used on its own during the test time. The tactile model aims to predict the 3d locations of both the hand and the object purely from the touch data by combining a predictive model and a contrastive learning module. This framework can reason about the interaction patterns from the tactile data, hallucinate the changes in the environment, estimate the uncertainty of the prediction, and generalize to unseen objects. We also provide detailed ablation studies regarding different system designs as well as visualizations of the predicted trajectories. This work takes a step on dynamics modeling in hand-object interactions from dense tactile sensing, which opens the door for future applications in activity learning, human-computer interactions, and imitation learning for robotics.

I. INTRODUCTION

When interacting with diverse objects using our hands, humans can feel and sense the contacts, enabling a strong intuitive understanding of the dynamics of the interactions (Figure 1a). This understanding allows us to predict how the

*Equal contribution. Q. Zhang and J. Yan are with Shanghai Jiao Tong University. Y. Li, Y. Luo, W. Shou, M. Foshey, J. B. Tenenbaum, W. Matusik, and A. Torralba are with the Computer Science and Artificial Intelligence Laboratory (CSAIL) at Massachusetts Institute of Technology.

environment will evolve and be influenced by these interactions, which is tremendously helpful for accomplishing daily manipulation tasks. To further our understanding of human activities and make progress in robotics, it is desirable to develop techniques that can model the dynamics of hand-object interactions, especially in highly dynamic scenarios. In recent years, there has been a growing amount of work aiming to understand and reconstruct hand-object interactions. Some of these works estimate the pose or reconstruct the shapes of both the objects and the hand [2], [3] , while others take a step further by modeling the contact when grasping an object [4], [5], [6]. However, most of the works focus on static grasps due to the reliance on indirect measurements such as thermal cameras to estimate the contact regions. The inability to provide an on-the-fly estimation of the contacts and the applied forces limits the use of these methods in more dynamic scenarios.

To directly sense the contacts, many researchers have built tactile sensors and explored how tactile sensing can recognize objects [7], [8], [9], classify materials [10], [11], [12], and how it can be coupled with the vision for representation learning [13], [14], [15]. Some prior works have tried to model the dynamics of the tactile interaction [16], [17], [18], yet most of them only use tactile sensing at the fingertip area and primarily focus on small motion. In this work, we aim to model the dynamics of hand-object interactions in highly dynamic tasks using a high-resolution tactile glove developed by Sundaram et al. [1] (Figure 1b). Specifically, we employ a

cross-modal learning framework that we collect synchronized RGBD images and tactile readings. During training, we use a visual processing pipeline to estimate the ground truth trajectories of both the object and the hand, which act as the labels to supervised the tactile model. During the test time, we only assume to have access to the tactile inputs and predict the corresponding system's states using the trained tactile model. The most desirable scenario would be to have a dexterous robotic hand performing the interactive tasks. However, although less accurate at repeating a specific action sequence, human capabilities are still far ahead of the current state-of-the-art robotic systems [19], [20], [21], [22]. Therefore, we ask a human subject to wear the glove to perform interactive activities in order *not* to limit ourselves from studying more dexterous and dynamic interactions with the objects. We consider four activities: (1) holding an object and waving it in the air, (2) throwing an object upwards and catch it (Figure 1c), (3) bouncing a ping-pong ball up and down using the racket, and (4) balancing an umbrella to prevent it from falling. Our goal is to predict the motion trajectory of the hand and the object purely based on the tactile readings (Figure 1d).

Our tactile model builds on top of a predictive model that takes the current state of the system, i.e., the positions and velocities of the object and the hand, and the tactile reading as input, then predicts the system state at the next time step. We add to the model an object encoder that produces object-centric embeddings calculated from a short sequence of interaction data that implicitly encodes the object-specific information like the mass and geometry. Experiments show that our model can reason about the interaction patterns from the tactile data, model the dynamics of the interactions, and quantify the uncertainty of the prediction. The use of the object encoder also allows an improved performance when generalizing to unseen objects. We also provide detailed ablation studies regarding the resolution of the tactile sensing and the visualizations of the predicted trajectories. This work moves in the direction of dynamics modeling in hand-object interactions from dense tactile sensing, which we hope can facilitate future studies and applications in multiple domains, including computer vision, graphics, and robotics.

II. RELATED WORK

1) Tactile sensing for perception and control: There has been a long history of hardware design for tactile sensing [23], [24], [25], [26]. Kappassov et al. [27] provide an excellent summary of tactile sensors for dexterous robot hands. Recently, Yuan et al. [28] propose to use a vision-based optical tactile sensor, which enables high-fidelity measurements of the contacts and object properties [29], [30]. Liu et al. [31], [32] build a tactile glove for studying object grasping and hand-object manipulation. Luo et al. [33] propose to build garments consist of conformal tactile textiles to learn human-environment interactions. In this paper, we use the tactile glove with a much higher resolution developed by Sundaram et al. [1]. The glove is flexible and deformable, giving us dense tactile sensing covering the palm area,

allowing us to record tactile information while performing various interactive tasks with diverse objects. With the help of these tactile sensors, previous researches improve the performance of various control tasks, ranging from basics object manipulation [34], [16], slip control [35], [36], to regrasping [37], [38] and others. This paper shares a similar spirit with [16] that learns to model the dynamics with tactile sensing, but mainly focus on hand-object interactions using a high-resolution tactile glove.

2) Modeling across multiple modalities: Frome et al. [39] seek to find the correlation between vision and language. Other researches also explored image captioning [40], [41] and learning image-text joint embeddings [42], [43], [44]. Finding the relationship between vision and sound is another interesting topic and has drawn increasing attention from the community [45], [46], [47]. Researchers have also proposed to connect vision and touch via cross-domain modeling [13], [48], [14], [15] or estimating 3D human poses from a tactile carpet by taking the outputs of computer vision models as supervisions [49]. Others have tried to combine tactile sensing with proprioception and kinesthetic information for object and shape recognition [50], [51]. In this paper, we also desire to establish a connection between touch and vision but focusing on the task of dynamics modeling using the newly-developed tactile glove [1].

3) Hand-object interactions: Recent advances in learning human-object interactions focus on estimating and tracking the hand in hand-object interaction scenarios [3], [52], [53], where most of the works take images as the primary sensing inputs [54], [55], [56]. Estimating and reasoning about the contacts is critical for building a better model of the inter-actions [4], [1]. Therefore, Brahmbhatt et al. [5], [6] use a thermal camera to label the contact regions during grasping. Some recent works [55], [57], [58] aim to infer contact force from visual images or take a step further by predicting the motion trajectories. Overall, prior methods rely on indirect measurements or weak supervision from a physics-based simulator and do not directly use real tactile data.

4) Contrastive learning: People have been using triplet loss for better representation learning [59], [60], while others have investigated the use of contrastive loss [61]. Recently, many papers use contrastive learning to get better self-supervised pre-trained features to boost the performance of downstream tasks [62], [63], [64]. Other researchers have been using contrastive learning to find better representa-tions via multi-task learning, e.g., Sun et al. [65] combine softmax and contrastive loss to help face recognition tasks. Zhang et al. [66] use triplet loss to get better features by embedding label structures. In this project, we use the contrastive loss to facilitate the model's generalization per-formance in multi-object scenarios, where we hope to learn an embedding space that contains object-specific information like the mass and the geometry.

III. METHOD

In this section, we present the details of our dynamics modeling framework. We first discuss some details of the
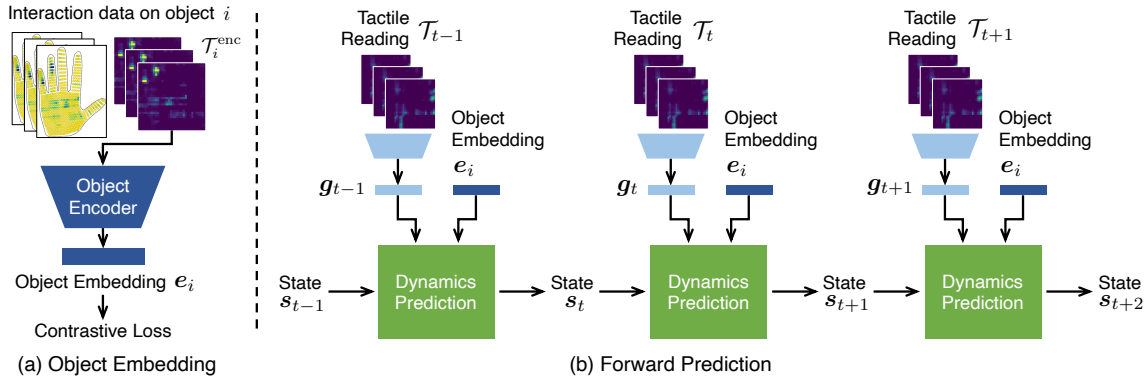
Fig. 2: **Dynamics Prediction Model.** (a) The object embedding e_i is derived by encoding a short period of interaction data into an implicit object-specific representation containing information like the mass and geometry. (b) Our model builds on top of a dynamics prediction model, which takes (i) the current state s_t, i.e., the positions and velocities of the object and the hand, (ii) the tactile embedding g_t, and (iii) the object embedding e_i as input, and then predicts the system state at the next step.

tactile glove and go through the overall structure of our dynamics module in Section III-B. We then describe how we quantify the uncertainty using pinball loss (Section III-C) and improve the generalization performance by extracting an object embedding using contrastive learning (Section III-D). In Section III-E, we describe our overall loss and how we measure the performance of the dynamics modeling.

A. Tactile Glove.

We use the tactile glove proposed by Sundaram et al. [1] to collect the hand-object interaction touch data (Figure 1b). There are about 680 sensors covering the palm area of the glove that can measure the applied forces. We record the tactile signal at each time step as an image of 32×32 resolution at a rate of around 15 Hz.

B. Dynamics Prediction Model.

The overall dynamics model is defined as the following:

$$\hat{s}_{t+1} = f(s_t, g_t, e_i), \qquad (1)$$

where s_t denotes the state of the system at time t, g_t indicates the tactile embedding derived from the current tactile input, and e_i is the object embedding that is shared across all time steps, containing contextual information like the mass and geometry of object i.

We represent the state of our system using the states of both the object and the hand, $s_t = (s_t^{\text{object}}, s_t^{\text{hand}})$, each of which is a concatenation of its position and velocity,

$$s_t^{\text{hand}} = (q_t^{\text{hand}}, \dot{q}_t^{\text{hand}}), \quad s_t^{\text{object}} = (q_t^{\text{object}}, \dot{q}_t^{\text{object}}), \qquad (2)$$

where $q_t = (x_t, y_t, z_t)$ indicates its 3D location.

To calculate the tactile embedding g_t, we take a small temporal window of M tactile frames as input, $\mathcal{T}_t = \{T_{t-M+1}, \ldots, T_{t-1}, T_t\}$, concatenate them to form a tensor of shape $M \times 32 \times 32$, and then feed it through an encoder (Figure 2b). We implement the encoder as a three-layer convolutional network with channel sizes of 32, 64, 128, individually. The dynamics prediction module (the green box in Figure 2b) is an MLP (Multi-Layer-Perceptron) that predicts the next state $\hat{s}_{t+1}$ by taking as input the concatenation of the tactile embedding g_t, the object embedding e_i, and the system state s_t at the last time step.

C. Pinball Loss for Uncertainty Estimation.

To quantify the uncertainty in state prediction, we use a pinball loss function [67] to predict both the mean value and the 10% and 90% quantiles. The intuition is that for a given set of inputs, the model estimates through a single forward pass both the expected future state and a cone of error within which the ground truth should lie 80% of the whole time.

More specifically, the outputs of the dynamics prediction module contain both the state $\hat{s}$ and the offset from the expected state, σ, such that $\hat{s} + \sigma$ is the 90% quantile of its distribution while $\hat{s} - \sigma$ is the 10% quantile. Assuming that the ground truth future state is s, the pinball loss is calculated as the following:

$$\mathcal{L}_{\text{pinball}}^{\tau} = \max(\tau \hat{p}_{\tau}, -(1-\tau)\hat{p}_{\tau}), \text{where}$$
$$\hat{p}_{\tau} = \begin{cases} s - (\hat{s} - \sigma), & \text{for } \tau \leq 0.5 \\ s - (\hat{s} + \sigma), & \text{otherwise.} \end{cases} \qquad (3)$$

We use the L1 regression loss and this pinball loss for both the predicted state and the quantiles at $\tau = 0.1$ and $\tau = 0.9$, making σ a measurement of the difference between the 10% and 90% quantiles and the expected value. The final loss for future state prediction is then:

$$\mathcal{L}_{\text{state}} = |s - \hat{s}| + \mathcal{L}_{\text{pinball}}^{0.1} + \mathcal{L}_{\text{pinball}}^{0.9}. \qquad (4)$$

D. Contrastive Learning for Object Embeddings.

To improve our model's generalization ability to unseen objects, we propose to encode a short period of interaction data into an implicit object embedding containing object-specific information like the mass and geometry. The intuition here is that when we humans have interacted with an object for a while using our hands, we can develop a better intuitive understanding of the object's dynamics and give better estimates of the object's physical properties like mass.

We subsample a short sequence from the interaction dataset of the object i and denote it as $\mathcal{T}_i^{\text{enc}}$. An encoder, implemented as a convolutional neural network, maps $\mathcal{T}_i^{\text{enc}}$ into its object embedding e_i (Figure 2a). When predicting multiple steps into the future using the dynamics prediction
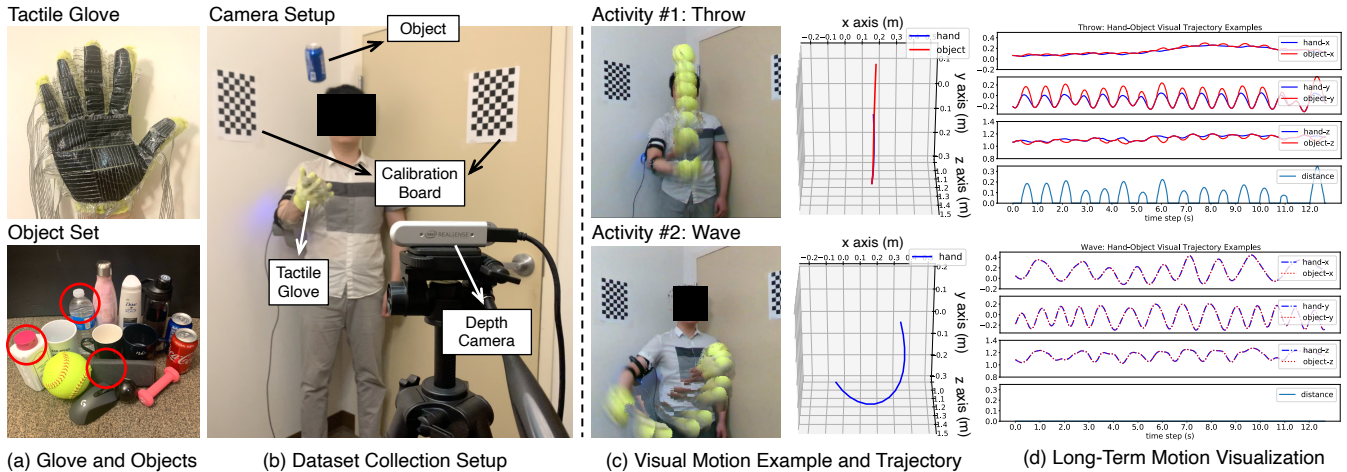
Fig. 3: **Data Collection Setup and Motion Examples.** (a) The tactile glove and the object set for performing "wave" and "throw" activities. Objects used for evaluating the model's generalization ability are marked using red circles (unseen during training). (b) A depth camera collects RGBD videos, which we use to project the hand and object's motion to 3D. (c) Here we show a visual example and the corresponding 3D trajectory for each activity. (d) shows long-term visualizations of typical motions in each axis and the distance between hand and object for the two activities.

module, the tactile embedding g_t is changing as it is derived from the current tactile inputs. In contrast, the object embedding e_i is kept the same throughout the prediction horizon (Figure 2b), which we hope can help the model disentangle the object-related information from the motion-related features, facilitating generalization to unseen objects.

Inspired by recent developments in contrastive representation learning [64], we introduce an auxiliary contrastive loss to help the object encoder learn more meaningful representations. We use the object embeddings calculated from the same object as the positive pair and embeddings from different objects as the negative pair. We then define the auxiliary contrastive loss as the following:

$$\mathcal{L}_{\text{obj}} = \mathbb{E}_e \left[\log \frac{\exp(\boldsymbol{e}^T \cdot \boldsymbol{e}^+)}{\exp(\boldsymbol{e}^T \cdot \boldsymbol{e}^+) + \Sigma_{j=1}^{N_{\text{sample}}-1} \exp(\boldsymbol{e}^T \cdot \boldsymbol{e}^-)} \right],$$
(5)

where $\boldsymbol{e}$ is the reference embedding, $\boldsymbol{e}^+$ and $\boldsymbol{e}^-$ are positive and negative embeddings respectively, and N_{sample} denotes the number of samples.

E. Overall Loss and Evaluation Metrics.

The overall loss of our framework is then a weighted sum of the future state prediction loss $\mathcal{L}_{\text{state}}$ and the contrastive loss for learning the object embeddings $\mathcal{L}_{\text{obj}}$:

$$\mathcal{L} = \mathcal{L}_{\text{state}} + \lambda_{\text{obj}}\mathcal{L}_{\text{obj}}.$$
(6)

We evaluate the performance of our dynamics modeling framework using two different metrics: (1) the L1 distance between the predicted state and the ground truth over different time steps, and (2) the Hausdorff distance [68] between the predicted and ground truth trajectories. Given the ground truth trajectory $A = \{\boldsymbol{s}_t, \boldsymbol{s}_{t+1}, ...\}$ and the predicted trajectory $B = \{\hat{\boldsymbol{s}}_t, \hat{\boldsymbol{s}}_{t+1}, ...\}$, the Hausdorff distance measures the similarity based on the shape of the trajectories and is defined as the following:

$$H = \max\{\max_{\boldsymbol{a}_i \in A} \min_{\boldsymbol{b}_j \in B} d(\boldsymbol{a}_i, \boldsymbol{b}_j), \max_{\boldsymbol{b}_i \in B} \min_{\boldsymbol{a}_j \in A} d(\boldsymbol{b}_i, \boldsymbol{a}_j)\}, \quad (7)$$

Dataset	Split	Object #	Train	Eval	All
Wave	Seen	12	166k	42k	208k
	Unseen	3	0	58k	58k
Throw	Seen	12	184k	46k	230k
	Unseen	3	0	61k	61k

TABLE I: **Dataset Statistics.** There are 12 seen objects and another 3 unseen objects (marked with red circle in Figure 3a) to evaluate the model's generalization ability. For the seen object dataset, 80% is kept for training while the remaining is for evaluation. For the unseen object, all of the data is reserved for evaluation.

where $d(\cdot, \cdot)$ denotes the Euclidean distance.

IV. EXPERIMENTS

A. Data Collection Setup.

The primary focus of this work is the dynamics modeling of hand-object interactions from tactile sensing. Given that there is still a huge performance gap between the state-of-the-art dexterous hand robotic system [19], [20], [21], [22] and human capabilities, we ask a human subject to wear the glove and perform various interaction tasks instead of using a robot system, which allows us to study more dexterous and dynamic interactions.

Figure 3b shows the setup for data collection, where a human subject faces a depth camera and performs various hand-object interaction activities. The RGBD camera and the tactile glove collect synchronized visual and tactile data with a framerate of 20 Hz and 15 Hz, respectively. To obtain the ground truth locations of both the hand and the object, we first segment the image according to the depth map and use the CSRT tracker [69] implemented in the OpenCV library to track their locations. We also calibrate the camera to ensure that we can interpret the ground truth labels in real physical units like meter and second. Compared with motion capture systems using markers on the object/hand, this RGBD-based caption system is simple, cheap, and easy to use. See the accompanying videos for visualizations of the tracking results.
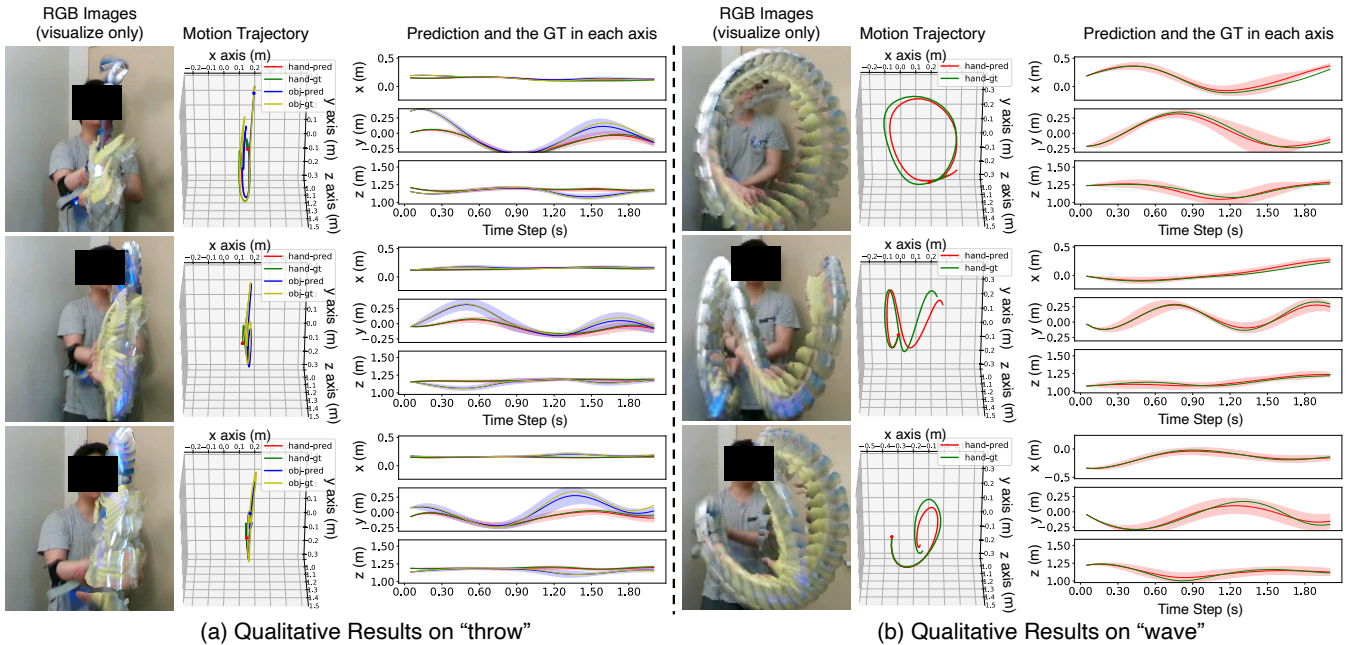
Fig. 4: **Qualitative Visualizations for "Throw" and "Wave".** (a) Visualization of the results in the "throw" activity. From left to right, we show the reference images (only used for visualization), the predicted and the ground truth trajectory in 3D, and the individual comparison in each axis. The shaded area indicates the estimated uncertainty from our model that corresponds to the 10% and 90% quantiles. (b) The same set of visualizations for the "wave" activity. Our model's predictions closely resemble what's happened in the real physical world.

We first consider the *Wave* activity, where the human subject holds an object in hand and wave it in the air. This activity is to verify the feasibility of the task: only when there exists a strong correlation between the tactile signal and the object's acceleration can we predict the motion trajectories purely from touch. This intuition on the correlation between different sensory modalities originates from Newton's second law of motion, which builds the connection between force (perceived by the tactile signal) and motion (the object's acceleration). Then we extend the system to more complicated activities where there can be large relative motion between the object and the hand:

- *Throw*: throw an object upward and catch it,
- *PingPong*: bounce a ping-pong ball using the racket,
- *Balance*: balance an umbrella to prevent it from falling.

Examples of the four activities are shown in Figure 3 and Figure 5 respectively. For *Wave* and *Throw*, we collect data for interacting with 15 different objects (Figure 3a), where we constantly change the in-hand position of the objects to increase the diversity of the dataset. We then split them into 12 seen and 3 unseen objects. Among the data for the seen objects, 80% of them are used for training, while the remaining 20% are for evaluation. The detailed statistics are shown in Table I.

B. Experimental Results on "Wave" and "Throw".

This section presents the experimental results on the *Wave* and *Throw* tasks. We first evaluate the model's performance by training separate models for each object. Then, we train a single model on all objects to test the model's multi-object generalization ability and show the benefits brought by the object embedding.

1) Results on Single Object.: We train a separate model for each object by minimizing the loss fucntion defined in Equation 6 for 50 epochs using Adam optimizer [70] and a learning rate of 0.0005.

We compare our model with the following two baselines: (1) Random: For each testing example, we randomly sample another trajectory from the dataset and measure their difference using the metrics discussed in Section III-E, which acts as a lower bound for the performance. (2) w/o Touch: In this baseline, we remove the tactile input and only use the state at the previous time step (s_t in Figure 2b) as the input to the dynamics prediction module. This baseline serves to show the benefit of the tactile information in dynamics modeling.

The left part of Table II illustrates the quantitative results. It is clear from the table that the touch information is critical for predicting the dynamics. We also include some qualitative results in Figure 4 for each activity by showing a side-by-side illustration of the original RGB image, the motion trajectories in 3D, and the estimated uncertainty in each axis. Our model can generate trajectories that closely match the ground truth, where the hand and object's movements consistently lie in our model's predicted confidence interval.

2) Results on Multi-Object Generalization: This section evaluates the model's ability to generalize across multiple objects. We use an object encoder to extract the implicit embeddings for each object that contain object-related information like the mass and the geometry (Figure 2a). To optimize our model, we first train the object encoder using contrastive loss introduced in Section III-D, where the positive and negative samples are from the same object and different objects separately. Each sample is a 20-step touch segment uniformly sampled from all touch data sequences. We then

Exp. type	Single object			Multi-object			
Setting	Average			Seen		Unseen	
Throw	Random	w/o Touch	Ours	w/o Embedding	Ours	w/o Embedding	Ours
L1@20	0.992	0.341	**0.227**	0.281	**0.254**	0.312	**0.265**
L1@50	1.005	0.547	**0.288**	0.323	**0.297**	0.391	**0.305**
Hausdorff	1.169	1.007	**0.659**	0.868	**0.775**	0.918	**0.842**
Wave	Random	w/o Touch	Ours	w/o Embedding	Ours	w/o Embedding	Ours
L1@20	0.308	0.186	**0.108**	0.135	**0.116**	0.154	**0.129**
L1@50	0.309	0.212	**0.132**	0.161	**0.137**	0.182	**0.154**
Hausdorff	1.671	1.498	**0.917**	1.154	**1.025**	1.308	**1.119**

TABLE II: **Quantitative Results for "throw" and "wave".** Here we assess our model's ability for predicting the dynamics of a single object and that of cross-object generalization. We evaluate the performance using the Hausdorff distance as well as the average L1 distance between the prediction and the ground truth when rolling 20 or 50 steps into the future. In both the "throw" and the "wave" activities, our model outperforms a random baseline and a baseline that does not use the touch information, which is in line with our intuition that touch plays a critical role in modeling the dynamics. The use of object embeddings also improves the performance, especially when encountering objects that are not seen during training.
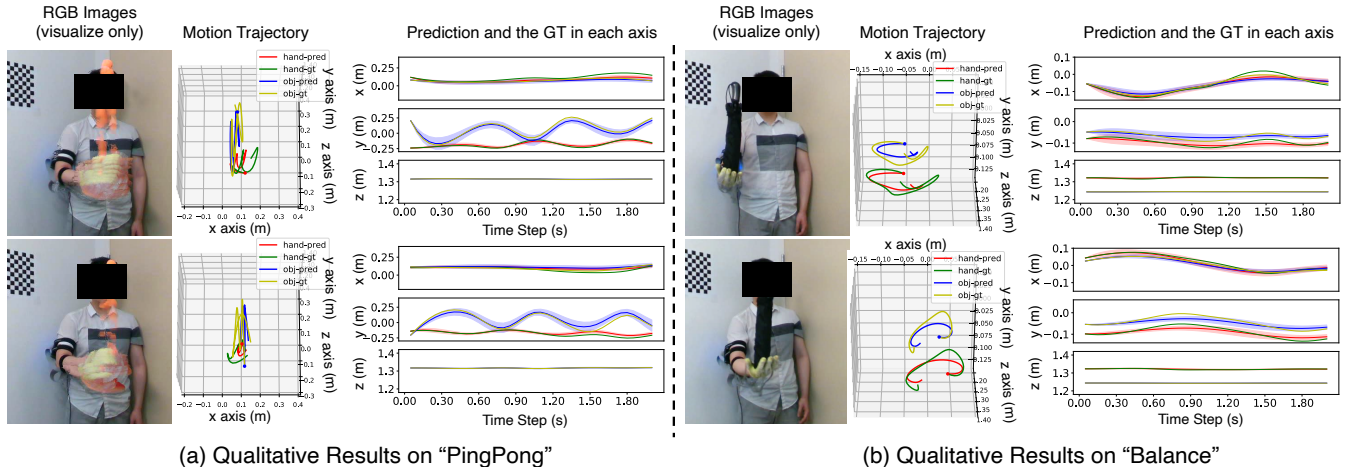

(a) Qualitative Results on "PingPong"

(b) Qualitative Results on "Balance"

Fig. 5: **Qualitative Visualizations for "Ping-Pong" and "Balance".** Similar to Figure 4, we show the reference images, the predicted and the ground truth 3D trajectory, and individual comparisons in each axis. The state of the object s_t^{object} is represented as the position and velocity of the ping-pong ball and the upper tip of the umbrella in these two activities, respectively. Again, our model's prediction closely matches the evolution of the actual physical systems.
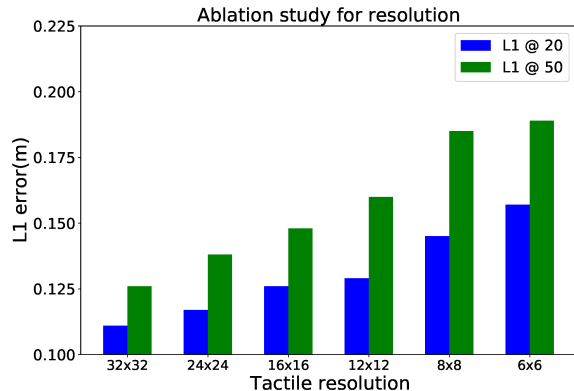

Fig. 6: **Ablation Study on the Resolution of the Tactile Glove.** The error bars show the L1 error of the prediction at 20 or 50 steps into the future when using tactile images of different resolutions, ranging from 32×32 to 6×6. Higher resolution can, in general, achieve better performance.

fix the object embedding for each object and concatenate it with the tactile embedding as the input feature for the subsequent computation (Figure 2b). During the test time, we choose the touch sequence from the first 20 steps and feed it into the object encoder to obtain the object embeddings.

We compare our model with a baseline that does not use the object embedding (w/o Embedding). The right half of Table II presents the quantitative results on cross-object generalization performance, which shows the benefit of the object embedding for improving the generalization ability.

3) Ablation Study on the Tactile Resolution: In this section, we aim to evaluate the importance of sensor resolution in dynamics modeling. We downsample the input tactile image from original 32×32 resolution to 24×24, 16×16, 12×12, 8×8, and 6×6 respectively, and evaluate the performance on the "Billiard" dataset from the "Waving" task. As shown in Figure 6, with the decrease of the effective resolution, the prediction performance drops notably, showing the importance of the sensor count in dynamics modeling.

C. Results on "PingPong" and "Balance".

We also evaluate our model's performance on the two other hand-object interaction activities: *PingPong* and *Balance*. The training and test numbers are 15k and 3k, respectively, for both of these two activities. Our model takes the tactile sequences as input and predicts the movements of the ping-pong ball and the umbrella. In *PingPong*, the state of the object s_t^{object} is represented as the position and velocity of the ping-pong ball, while in *Balance*, we use the position and velocity of the umbrella's upper tip as the object state. The quantitative results and the qualitative results are shown

PingPong	Random	w/o Touch	Ours
L1@20	0.205	0.164	**0.092**
L1@50	0.207	0.185	**0.109**
Hausdorff	0.725	0.293	**0.162**
Balance	Random	w/o Touch	Ours
L1@20	0.183	0.139	**0.076**
L1@50	0.187	0.175	**0.103**
Hausdorff	0.835	0.453	**0.267**

TABLE III: Quantitative Results for "PingPong" and "Balance". Our method outperforms two baselines on all the metrics, demonstrating the effectiveness of our dynamics prediction model and the critical role tactile sensing played in the process.

in Table III and Figure 5, respectively.

V. CONCLUSION

In this paper, we study the problem of dynamics modeling in hand-object interactions using tactile sensing, where a human subject acts as a proxy for a robotic system to investigate more dexterous and dynamic interactive activities. The human subject wears a high-resolution tactile glove to perform four different interactive tasks with a range of diverse objects. We build a model that predicts how the environment is expected to evolve with time by only taking the current state and the subsequent tactile information as input. We further augment the model using an object-centric encoder to facilitate multi-object generalization. Experiments show that our model can predict the dynamics of the interactions, quantify the uncertainty of the prediction, and generalize to unseen objects. Through this work, we hope to draw people's attention and inspire future investigations to build better perception systems and intuitive models of the environment via the lens of different sensing modalities.

In the future, we desire to predict not only the position but also the orientation of the object, where we plan to combine the proposed framework with other sources of information like IMUs, or human (or robot) joint states (e.g., positions, torques), which we believe would improve and scale up the framework for more complicated applications and tasks.

REFERENCES

[1] S. Sundaram, P. Kellnhofer, Y. Li, J.-Y. Zhu, A. Torralba, and W. Matusik, "Learning the signatures of the human grasp using a scalable tactile glove," *Nature*, vol. 569, no. 7758, pp. 698–702, 2019.

[2] O. Glauser, S. Wu, D. Panozzo, O. Hilliges, and O. Sorkine-Hornung, "Interactive hand pose estimation using a stretch-sensing soft glove," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1–15, 2019.

[3] Y. Hasson, G. Varol, D. Tzionas, I. Kalevatykh, M. J. Black, I. Laptev, and C. Schmid, "Learning joint reconstruction of hands and manipulated objects," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 807–11 816.

[4] K. Bernardin, K. Ogawara, K. Ikeuchi, and R. Dillmann, "A sensor fusion approach for recognizing continuous human grasping sequences using hidden markov models," *IEEE Transactions on Robotics*, vol. 21, no. 1, pp. 47–57, 2005.

[5] S. Brahmbhatt, C. Ham, C. C. Kemp, and J. Hays, "Contactdb: Analyzing and predicting grasp contact via thermal imaging," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8709–8719.

[6] S. Brahmbhatt, C. Tang, C. D. Twigg, C. C. Kemp, and J. Hays, "Contactpose: A dataset of grasps with object contact and hand pose," *arXiv preprint arXiv:2007.09545*, 2020.

[7] J. M. Gandarias, A. J. Garcia-Cerezo, and J. M. Gomez-de Gabriel, "Cnn-based methods for object recognition with high-resolution tactile sensors," *IEEE Sensors Journal*, vol. 19, no. 16, pp. 6872–6882, 2019.

[8] T. Corradi, P. Hall, and P. Iravani, "Bayesian tactile object recognition: Learning and recognising objects using a new inexpensive tactile sensor," in *2015 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2015, pp. 3909–3914.

[9] M. Bauza, O. Canal, and A. Rodriguez, "Tactile mapping and localization from high-resolution tactile imprints," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 3811–3817.

[10] Z. Erickson, S. Chernova, and C. C. Kemp, "Semi-supervised haptic material recognition for robots using generative adversarial networks," *arXiv preprint arXiv:1707.02796*, 2017.

[11] W. Zheng, B. Wang, H. Liu, X. Wang, Y. Li, and C. Zhang, "Bioinspired magnetostrictive tactile sensor for surface material recognition," *IEEE Transactions on Magnetics*, vol. 55, no. 7, pp. 1–7, 2019.

[12] Y. Wang, J. Chen, and D. Mei, "Recognition of surface texture with wearable tactile sensor array: A pilot study," *Sensors and Actuators A: Physical*, p. 111972, 2020.

[13] W. Yuan, S. Wang, S. Dong, and E. 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 (CVPR)*, July 2017.

[14] Y. Li, J.-Y. Zhu, R. Tedrake, and A. Torralba, "Connecting touch and vision via cross-modal prediction," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[15] J.-T. Lee, D. Bollegala, and S. Luo, ""touching to see" and "seeing to feel": Robotic cross-modal sensory data generation for visual-tactile perception," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 4276–4282.

[16] S. Tian, F. Ebert, D. Jayaraman, M. Mudigonda, C. Finn, R. Calandra, and S. Levine, "Manipulation by feel: Touch-based control with deep predictive models," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 818–824.

[17] Y. She, S. Wang, S. Dong, N. Sunil, A. Rodriguez, and E. Adelson, "Cable manipulation with a tactile-reactive gripper," *arXiv preprint arXiv:1910.02860*, 2019.

[18] Y. S. Narang, K. Van Wyk, A. Mousavian, and D. Fox, "Interpreting and predicting tactile signals via a physics-based and data-driven framework," *arXiv preprint arXiv:2006.03777*, 2020.

[19] S. Kim, A. Shukla, and A. Billard, "Catching objects in flight," *IEEE Transactions on Robotics*, vol. 30, no. 5, pp. 1049–1065, 2014.

[20] I. Akkaya, M. Andrychowicz, M. Chociej, M. Litwin, B. McGrew, A. Petron, A. Paino, M. Plappert, G. Powell, R. Ribas, *et al.*, "Solving rubik's cube with a robot hand," *arXiv preprint arXiv:1910.07113*, 2019.

[21] O. M. Andrychowicz, B. Baker, M. Chociej, R. Jozefowicz, B. McGrew, J. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray, *et al.*, "Learning dexterous in-hand manipulation," *The International Journal of Robotics Research*, vol. 39, no. 1, pp. 3–20, 2020.

[22] A. Nagabandi, K. Konolige, S. Levine, and V. Kumar, "Deep dynamics models for learning dexterous manipulation," in *Conference on Robot Learning*. PMLR, 2020, pp. 1101–1112.

[23] M. R. Cutkosky, R. D. Howe, and W. R. Provancher, "Force and tactile sensors." *Springer Handbook of Robotics*, vol. 100, pp. 455–476, 2008.

[24] S. J. Lederman and R. L. Klatzky, "Haptic perception: A tutorial," *Attention, Perception, & Psychophysics*, vol. 71, no. 7, pp. 1439–1459, 2009.

[25] M. K. Johnson and E. H. Adelson, "Retrographic sensing for the measurement of surface texture and shape," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 1070–1077.

[26] M. K. Johnson, F. Cole, A. Raj, and E. H. Adelson, "Microgeometry capture using an elastomeric sensor," *ACM Transactions on Graphics (TOG)*, vol. 30, no. 4, pp. 1–8, 2011.

[27] Z. Kappassov, J.-A. Corrales, and V. Perdereau, "Tactile sensing in dexterous robot hands," *Robotics and Autonomous Systems*, vol. 74, pp. 195–220, 2015.

[28] W. Yuan, S. Dong, and E. H. Adelson, "Gelsight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, vol. 17, no. 12, p. 2762, 2017.

[29] W. Yuan, M. A. Srinivasan, and E. H. Adelson, "Estimating object hardness with a gelsight touch sensor," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 208–215.

[30] W. Yuan, C. Zhu, A. Owens, M. A. Srinivasan, and E. H. Adelson, "Shape-independent hardness estimation using deep learning and a

gelsight tactile sensor," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 951–958.

[31] H. Liu, X. Xie, M. Millar, M. Edmonds, F. Gao, Y. Zhu, V. J. Santos, B. Rothrock, and S.-C. Zhu, "A glove-based system for studying hand-object manipulation via joint pose and force sensing," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 6617–6624.

[32] H. Liu, Z. Zhang, X. Xie, Y. Zhu, Y. Liu, Y. Wang, and S.-C. Zhu, "High-fidelity grasping in virtual reality using a glove-based system," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 5180–5186.

[33] Y. Luo, Y. Li, P. Sharma, W. Shou, K. Wu, M. Foshey, B. Li, T. Palacios, A. Torralba, and W. Matusik, "Learning human–environment interactions using conformal tactile textiles," *Nature Electronics*, vol. 4, no. 3, pp. 193–201, 2021.

[34] M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 8943–8950.

[35] F. Veiga, J. Peters, and T. Hermans, "Grip stabilization of novel objects using slip prediction," *IEEE transactions on haptics*, vol. 11, no. 4, pp. 531–542, 2018.

[36] S. Dong, D. Ma, E. Donlon, and A. Rodriguez, "Maintaining grasps within slipping bounds by monitoring incipient slip," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 3818–3824.

[37] Y. Chebotar, K. Hausman, Z. Su, G. S. Sukhatme, and S. Schaal, "Self-supervised regrasping using spatio-temporal tactile features and reinforcement learning," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 1960–1966.

[38] F. R. Hogan, M. Bauza, O. Canal, E. Donlon, and A. Rodriguez, "Tactile regrasp: Grasp adjustments via simulated tactile transformations," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 2963–2970.

[39] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," in *Advances in neural information processing systems*, 2013, pp. 2121–2129.

[40] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.

[41] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[42] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean, "Zero-shot learning by convex combination of semantic embeddings," *arXiv preprint arXiv:1312.5650*, 2013.

[43] M. Otani, Y. Nakashima, E. Rahtu, J. Heikkilä, and N. Yokoya, "Learning joint representations of videos and sentences with web image search," in *European Conference on Computer Vision*. Springer, 2016, pp. 651–667.

[44] Y. Aytar, L. Castrejon, C. Vondrick, H. Pirsiavash, and A. Torralba, "Cross-modal scene networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 10, pp. 2303–2314, 2017.

[45] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba, "Ambient sound provides supervision for visual learning," in *European conference on computer vision*. Springer, 2016, pp. 801–816.

[46] C. Gan, D. Huang, P. Chen, J. B. Tenenbaum, and A. Torralba, "Foley music: Learning to generate music from videos," *arXiv preprint arXiv:2007.10984*, 2020.

[47] H. Zhao, C. Gan, W.-C. Ma, and A. Torralba, "The sound of motions," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 1735–1744.

[48] P. Falco, S. Lu, A. Cirillo, C. Natale, S. Pirozzi, and D. Lee, "Cross-modal visuo-tactile object recognition using robotic active exploration," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 5273–5280.

[49] Y. Luo, Y. Li, M. Foshey, W. Shou, P. Sharma, T. Palacios, A. Torralba, and W. Matusik, "Intelligent carpet: Inferring 3d human pose from tactile signals," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 255–11 265.

[50] S. Luo, W. Mou, K. Althoefer, and H. Liu, "iclap: Shape recognition by combining proprioception and touch sensing," *Autonomous Robots*, vol. 43, no. 4, pp. 993–1004, 2019.

[51] F. Pastor, J. García-González, J. M. Gandarias, D. Medina, P. Closas, A. J. García-Cerezo, and J. M. Gómez-de Gabriel, "Bayesian and neural inference on lstm-based object recognition from tactile and kinesthetic information," *IEEE Robotics and Automation Letters*, vol. 6, no. 1, pp. 231–238, 2020.

[52] I. Oikonomidis, N. Kyriazis, and A. A. Argyros, "Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 2088–2095.

[53] S. Sridhar, F. Mueller, M. Zollhöfer, D. Casas, A. Oulasvirta, and C. Theobalt, "Real-time joint tracking of a hand manipulating an object from rgb-d input," in *European Conference on Computer Vision*. Springer, 2016, pp. 294–310.

[54] N. Kyriazis and A. Argyros, "Scalable 3d tracking of multiple interacting objects," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3430–3437.

[55] T.-H. Pham, A. Kheddar, A. Qammaz, and A. A. Argyros, "Towards force sensing from vision: Observing hand-object interactions to infer manipulation forces," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2810–2819.

[56] D. Tzionas, L. Ballan, A. Srikantha, P. Aponte, M. Pollefeys, and J. Gall, "Capturing hands in action using discriminative salient points and physics simulation," *International Journal of Computer Vision*, vol. 118, no. 2, pp. 172–193, 2016.

[57] T.-H. Pham, N. Kyriazis, A. A. Argyros, and A. Kheddar, "Hand-object contact force estimation from markerless visual tracking," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2883–2896, 2017.

[58] K. Ehsani, S. Tulsiani, S. Gupta, A. Farhadi, and A. Gupta, "Use the force, luke! learning to predict physical forces by simulating effects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 224–233.

[59] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 539–546.

[60] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[61] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1735–1742.

[62] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, pp. 297–304.

[63] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3733–3742.

[64] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[65] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Advances in neural information processing systems*, 2014, pp. 1988–1996.

[66] X. Zhang, F. Zhou, Y. Lin, and S. Zhang, "Embedding label structures for fine-grained feature representation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1114–1123.

[67] R. Koenker and K. F. Hallock, "Quantile regression," *Journal of economic perspectives*, vol. 15, no. 4, pp. 143–156, 2001.

[68] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the hausdorff distance," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 15, no. 9, pp. 850–863, 1993.

[69] A. Lukežič, T. Voj'iř, L. Čehovin Zajc, J. Matas, and M. Kristan, "Discriminative correlation filter tracker with channel and spatial reliability," *International Journal of Computer Vision*, 2018.

[70] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.