ELEPHANT: RECOGNIZE AND RECONSTRUCT THE WORLD BLINDLY

Anonymous authors
Paper under double-blind review



Figure 1: **Overview of Elephant.** From a few spatially localized tactile observations, Elephant jointly infers object category and complete 3D shape. We study this setting in both simulation and real-world robot experiments, and evaluate how object understanding improves as additional contacts are observed.

ABSTRACT

Understanding an object's geometry and semantic identity through physical contact is an important capability for robots operating with limited visual information. Yet each tactile observation reveals only a small region of the surface, leaving open how much object-level understanding can be achieved from just a few contacts. We introduce *Elephant*, a framework for joint object reconstruction and recognition from sparse tactile observations, without visual input at inference time. Elephant uses sensor poses to express local observations in a common spatial frame and aggregates evidence across grasps to infer global geometry and semantic category. We also introduce a real-world tactile dataset collected using UMI-based grippers equipped with optical Daimon and magnetic eFlesh sensors, pairing tactile measurements with sensor poses and object geometry. The curated Daimon subset comprises 52 objects from 3 families, while the eFlesh study covers 6 semantic categories composed of 18 instances. These real-world observations are complemented by simulated tactile data from 279 object instances spanning 43 semantic categories. On real Daimon data, Elephant achieves a CD-L1 of 6.19 mm with ten grasps. In a separate six-category eFlesh study with six held-out object instances, it achieves 94.4% recognition accuracy over 18 sampled inputs at 20 tactile episodes, alongside a reconstruction F-score of 56.2%, compared with 51.8% for the strongest evaluated reconstruction baseline. These results highlight the potential of spatially localized tactile observations for recovering global object shape and semantic identity across optical and magnetic sensing technologies.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Touch provides direct information about an object's contacted surface, offering a local source of evidence for understanding and interacting with the physical world. A substantial body of research investigates touch in combination with other sensing modalities, particularly vision, for perception and manipulation. For example, visuo-tactile reconstruction methods demonstrate how contact measurements complement visual estimates of object geometry (Smith et al., 2020; Suresh et al., 2024). These advances motivate a complementary question: how much can a few contacts reveal about an object's shape and category? We investigate this question through joint geometric reconstruction and semantic classification from sparse tactile observations, without visual input at inference time.

Prior work, including TouchSDF (Comi et al., 2024) and Touch2Shape (Wang et al., 2025), demonstrates tactile-only shape reconstruction, primarily focusing on geometry. We extend this investigation to joint reconstruction and semantic recognition across a broader set of categories than their six-category ShapeNet (Chang et al., 2015) evaluations.

We present *Elephant*, which uses sensor poses to align local contact geometry in a common coordinate frame and fuse evidence across grasps, predicting a complete surface point cloud and an object category without object images or visual point clouds at inference time.

Our contributions are:

- A real-world tactile dataset collected with UMI-based grippers equipped with Daimon or eFlesh (Pattabiraman et al., 2025) sensors, pairing tactile observations with sensor poses, object geometry, and semantic labels.
- Elephant, a framework combining sensor-specific encoders with a shared geometric and semantic inference backbone for object-level prediction from sparse contacts.
- An empirical study of reconstruction and recognition over one to ten grasps in simulation and on real sensors, characterizing how additional contacts improve predictions and what ambiguities persist under sparse coverage (Figure 1).

2 RELATED WORK

	Core Capability		Inference Setting		Transfer & Deployment		
Method	**Complete Object Reconstruction**	**Object Recognition**	**Aux. Object Cue**	**Few-Touch Inference**	**Real Hardware**	**Sim-to -Real**	**Cross Sensor**
Object shape and identity inference							
Smith et al. (2020)	✓	✗	Opt. RGB	✓	✗	✗	✗
TAPCNet (Liu et al., 2025)	✓	✗	Partial PC	✓	✗	✗	✗
TouchSDF (Comi et al., 2024)	✓	✗	None	✓	✓	✓	✗
Touch2Shape (Wang et al., 2025)	✓	✗	Opt. RGB	✓	✗	✗	✗
TouchAnything (Gu et al., 2026)	✓	✗	Class Text	✓	✓	✗	✗
Zheng et al. (2026)	✓	✓	None	✗	✗	✗	✗
Tactile representation and transfer							
SITR (Gupta et al., 2025)	✗	✓	None	✗	✓	✓	✓
AnyTouch (Feng et al., 2025)	✗	✗	–	✗	✓	✗	✓
TactSpace (Joarder et al., 2026)	✗	✗	–	✗	✓	✓	✗
TactX (Park et al., 2026)	✗	✓	None	✗	✓	✗	✓
Elephant (Ours)	✓	✓	None	✓	✓	✓	✓

Table 1: Comparison with representative tactile object inference and transfer methods. *Aux. Object Cue* denotes additional object-specific non-tactile input, and *Few-Touch Inference* denotes object-level inference from a fixed, small set of spatially sparse contacts. *Cross-Sensor* refers to transfer across distinct physical tactile sensor designs. ✓/✗ indicate whether a capability or setting is explicitly studied; – denotes not applicable.

Prior work has studied tactile perception from complementary perspectives: reconstructing object geometry from contact observations, transferring tactile knowledge beyond simulation, and adapting representations across sensing hardware. Table 1 summarizes representative methods along these axes and highlights the setting studied by Elephant.

2.1 Tactile Object Reconstruction and Recognition

Touch provides high-fidelity local surface evidence but leaves much of an object's global geometry unobserved (Smith et al., 2020; Suresh et al., 2022). Visuo-tactile reconstruction methods therefore combine local contact measurements with visual observations or partial geometry to recover or refine global object shape (Smith et al., 2020; Rustler et al., 2022; Suresh et al., 2022; Liu et al., 2025; Fang et al., 2025). Other approaches recover global shape from tactile observations using learned geometric representations or generative priors, with several additionally selecting subsequent contacts through active exploration (Smith et al., 2021; Comi et al., 2024; Murali et al., 2023; Wang et al., 2025; Zheng et al., 2026; Gu et al., 2026). A related line of work uses touch for object recognition and identification, often accumulating evidence over sequential contacts or actively selecting informative interactions (Zhang et al., 2017; Xu et al., 2022; Pai et al., 2023; Zheng et al., 2026). Together, these works establish touch as a cue for both global geometry and object identity, across settings that variously use auxiliary visual or partial geometry, active acquisition, or explicit semantic and geometric priors. Our work instead considers a fixed and highly limited set of tactile observations, jointly inferring complete object geometry and semantic category without auxiliary partial geometry or a class-level description at inference time.

2.2 Sim-to-Real Tactile Transfer

Simulation enables scalable tactile supervision, yet reproducing real contact mechanics and sensor responses remains challenging even with high-fidelity simulators (Si & Yuan, 2022). Synthetic-data approaches transfer simulation-trained models for force estimation and object classification (Sferrazza et al., 2020; Zhong et al., 2025), while domain-adaptation methods bridge simulated and real observations through image translation, appearance adaptation, or learned mappings (Church et al., 2022; Jianu et al., 2022; Chen et al., 2022; Azulay et al., 2024). For contact-rich control, improved simulation, learned tactile representations, and policy learning enable transfer to physical robots (Chen et al., 2024; Zhao et al., 2024; Wu et al., 2024; Akinola et al., 2025). Recent methods further learn sensor- or modality-invariant embeddings for zero-shot sim-to-real transfer (Gupta et al., 2025; Joarder et al., 2026). These studies span tactile perception, classification, contact geometry estimation, and manipulation. Our work investigate a complementary object-level setting: jointly recovering complete object geometry and semantic category using simulation-trained tactile knowledge and a fixed, highly limited set of spatially sparse real-world contacts.

2.3 Cross-Sensor Tactile Representation and Adaptation

Tactile sensors produce heterogeneous observations due to differences in hardware and transduction mechanisms. Hardware-oriented methods support reuse across instances of the same sensor design, while adaptation and translation methods transfer models or signals between related platforms (Bhirangi et al., 2025; Jing & Qian, 2025; Rodriguez et al., 2025). Unified models learn transferable features through sensor-specific conditioning, shared backbones, or aligned multi-sensor observations (Yang et al., 2024; Zhao et al., 2025; Feng et al., 2025). Recent approaches further align non-visual, cross-modality, and cross-embodiment observations in shared latent spaces for perceptual or policy transfer (Hou et al., 2025; Park et al., 2026). We study a complementary object-level setting: adapting sensor-specific encoders to a shared geometric and semantic inference backbone to recover complete object shape and category from a few spatially sparse contacts.

3 Method

3.1 Problem Formulation

Given K tactile observations $\mathcal{O}_K = \{\mathcal{O}_i\}_{i=1}^{K}$, Elephant predicts a complete surface point cloud $\hat{\mathcal{P}}$ and a distribution $\boldsymbol{\pi} \in [0,1]^C$ over C known semantic categories. Each observation contains tactile measurements and sensor poses for the contacting fingers. A *contact unit* comprises measurements from one finger in one observation; a bilateral observation contributes two units. The budget K counts observations, rather than fingers or points, with sensor-specific sampling protocols defined in the experiments.

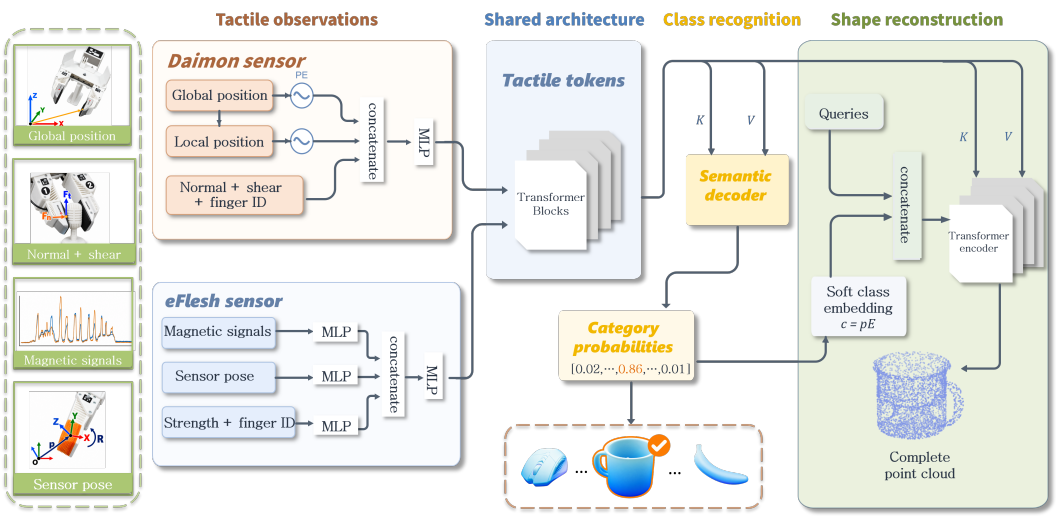

Figure 2: Elephant architecture. Tactile features support semantic recognition and condition shape reconstruction through both encoded geometry and the predicted category distribution.

We assume known sensor poses, finger identities, and calibration in a common object frame. For real data, this frame is established by offline RGB-D/CAD alignment. The network receives localized tactile features, without reference images, visual point clouds, complete-shape targets, or ground-truth categories as direct inputs; it does not estimate object pose. As shown in Figure 2, Elephant encodes geometry within and across contacts, predicts a category distribution, and uses that distribution together with tactile features to reconstruct the complete shape.

3.2 Spatially Aligned Tactile Representation

For the Daimon surface-point input, a sensor-frame point is mapped to the object frame by

$$\tilde{\mathbf{p}}^{O} = \mathbf{T}_{O\leftarrow C}\,\mathbf{T}_{C\leftarrow M}\,\mathbf{T}_{M\leftarrow S}\,\tilde{\mathbf{p}}^{S}, \tag{1}$$

where tildes denote homogeneous coordinates, and S, M, C, and O denote the sensor, marker, camera, and object frames. As illustrated in Figure 3, the transforms combine sensor-to-marker calibration, marker tracking in the camera frame, and CAD-to-camera alignment.

We center inputs and targets at the observed tactile centroid and divide by a fixed scale. Predictions are mapped back by inverting this normalization.

For a normalized point $\mathbf{p}_i$ in contact unit u, the input token is

$$\mathbf{h}_i = \mathrm{MLP}([\gamma(\mathbf{p}_i),\, \gamma(\mathbf{p}_i - \boldsymbol{\mu}_u),\, \mathbf{n}_i,\, \mathbf{s}_i,\, q_i,\, \mathbf{e}_{f_i}])\,, \tag{2}$$

where γ is Fourier encoding, $\boldsymbol{\mu}_u$ is the unit centroid, and $\mathbf{n}_i$, $\mathbf{s}_i$, q_i, and $\mathbf{e}_{f_i}$ represent surface normal, shear, contact confidence, and finger identity. Global and unit-relative positions encode the contact arrangement and local patch geometry, respectively. Contact-unit membership is retained throughout the geometric encoding.

3.3 Hierarchical Multi-Touch Encoding and Recognition

Intra-contact encoding. We construct spatial k-nearest-neighbor graphs within each contact unit and keep their neighborhoods fixed across layers. Each layer updates point features as

$$\mathbf{h}_i' = \mathbf{h}_i + \max_{j\in\mathcal{N}_k(i)} \phi\big([\bar{\mathbf{h}}_i,\, \bar{\mathbf{h}}_j - \bar{\mathbf{h}}_i,\, \mathbf{p}_j - \mathbf{p}_i]\big)\,, \tag{3}$$

where $\bar{\mathbf{h}}$ denotes layer normalization, ϕ is a layer-specific MLP, and the maximum is channelwise. These local graphs encode surface structure without connecting spatially separated patches.

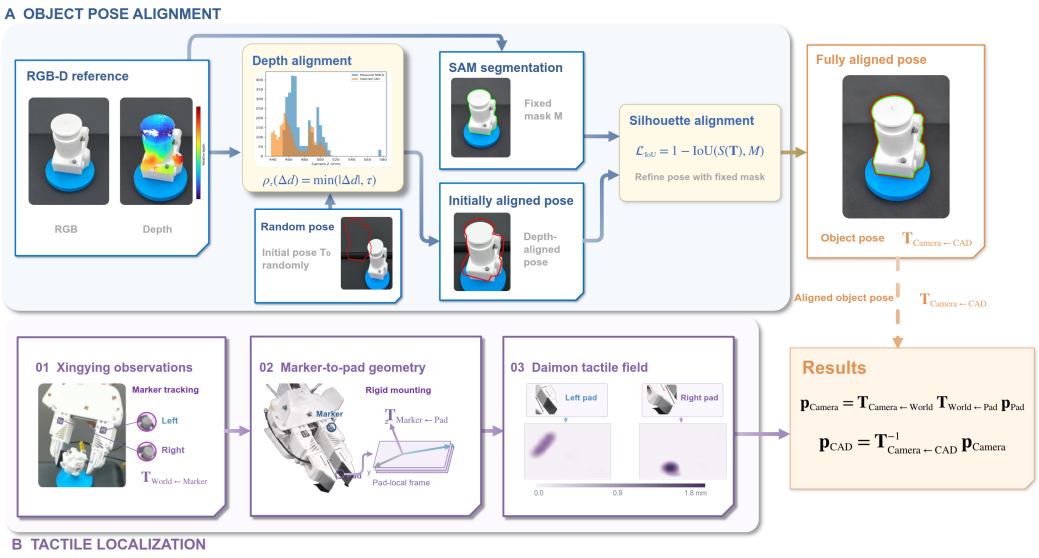

Figure 3: Spatial alignment of tactile observations and the CAD surface. Reference RGB-D data establish the object pose, while tracked sensor poses and calibration localize tactile patches in the object frame.

Inter-contact attention. Each unit is summarized by its mean feature, centroid, and normalized mean normal. For units (u, v), a relation vector $\mathbf{r}_{uv}$ concatenates their centroid displacement, distance, normal difference, and normal dot product. Attention head a uses this geometry as an additive bias:

$$\alpha_{uv}^{(a)} = \mathrm{softmax}_v\left(\frac{(\mathbf{q}_u^{(a)})^\top \mathbf{k}_v^{(a)}}{\sqrt{d_a}} + \psi_a(\mathbf{r}_{uv})\right), \tag{4}$$

where queries and keys are projected from normalized unit features, d_a is the head dimension, and ψ_a is an MLP. The attention-weighted value message is projected and added residually to every point token in its unit, combining cross-contact context with pointwise geometry.

Global aggregation and recognition. A Transformer encodes the enriched point tokens into $\mathbf{H}$ without sequential touch-order embeddings. Validity masks exclude padding from graph construction, attention, and pooling. A learnable classification query attends to $\mathbf{H}$, followed by an MLP and softmax to predict $\boldsymbol{\pi}$. The same pointwise features $\mathbf{H}$ provide geometric context for reconstruction.

3.4 Semantic-Conditioned Shape Reconstruction

To guide completion of unobserved regions, we condition reconstruction on the predicted category distribution through learnable category embeddings:

$$\mathbf{c} = \sum_{r=1}^{C} \pi_r \mathbf{E}_r. \tag{5}$$

We add $\mathbf{c}$ to each learnable reconstruction query, preserving soft category hypotheses rather than selecting a single label. Each decoder layer applies cross-attention to $\mathbf{H}$, self-attention among queries, and a feed-forward transformation. Each of the 512 final queries predicts a coarse point as an offset from a learnable spatial template; a second head predicts four offsets around each coarse point, yielding a 2,048-point surface. Predicted category probabilities provide conditioning during both training and inference.

3.5 Training Objectives

We jointly optimize reconstruction and recognition with $\mathcal{L} = \mathcal{L}_{\text{coarse}} + \mathcal{L}_{\text{fine}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}}$. Both reconstruction terms use symmetric Chamfer distance, averaging the two directional mean nearest-neighbor distances with unsquared Euclidean distance. The fine target is the complete surface; the coarse target is its 512-point farthest-point subsample. Classification uses cross-entropy, weighted by λ_{cls}. Both tasks update the shared encoder, and reconstruction gradients also reach the category embeddings and classification branch through the soft semantic conditioning.

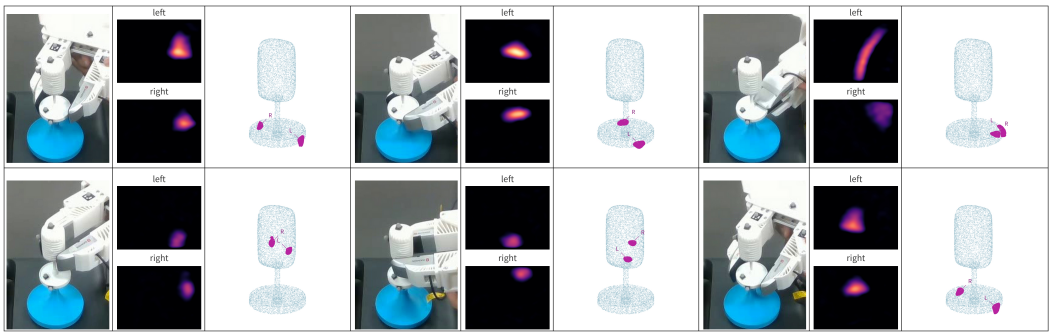

Figure 4: Visualization of tactile data collection. Each example shows the depth camera's view, left/right tactile observations, and the corresponding contact locations on the object point cloud.

4 Experiments

Our experiments address three questions: (1) How do object reconstruction and category recognition vary with the tactile observation budget? (2) How well does a simulation-trained model transfer to real Daimon observations? (3) Can the same inference architecture support reconstruction and recognition from sparse magnetic eFlesh measurements using a sensor-specific encoder?

4.1 Datasets and Experimental Setup

Simulation Datasets. Our simulation dataset contains 279 object instances spanning 43 semantic categories, drawn from ShapeNet (Chang et al., 2015), YCB (Calli et al., 2015), Google Scanned Objects (Downs et al., 2022), and Objaverse (Deitke et al., 2023) (Figure 5). We use 207 instances from 41 categories for training and validation, and hold out 59 additional instances from these categories for evaluation. Another 13 instances from two unseen categories are evaluated only for reconstruction, as these categories are outside the classifier's training vocabulary.

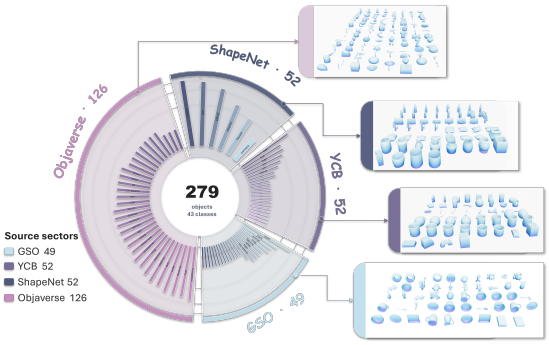

Figure 5: Object sources and category distribution in the simulation dataset, comprising 279 object instances across 43 semantic categories.

Each object is normalized and imported into MuJoCo, where a pair of opposing tactile pads is randomly positioned and closed around the object to generate grasp-based tactile observations. For each grasp, we record local tactile observations and sensor poses, while retaining the complete object surface as reconstruction supervision.

Real-world Tactile Datasets. We collect real-world tactile data using UMI-based grippers equipped with optical Daimon and magnetic eFlesh sensors. Figure 4 visualizes representative

recordings, including the camera view, bilateral tactile observations, and their corresponding contact locations on the object surface.

The Daimon dataset contains 52 physical objects across 674 recordings, with seven objects held out entirely for testing. Notably, seven of the Daimon assets are components of the UMI platform itself, providing a small but deliberate testbed for tactile self-component recognition.

The eFlesh dataset contains 18 object instances across six semantic categories, with one complete instance per category held out from training. The two datasets are collected independently and use sensor-specific tactile measurements with corresponding poses, object geometry, and semantic labels. Figure 6 summarizes the Daimon object collection.

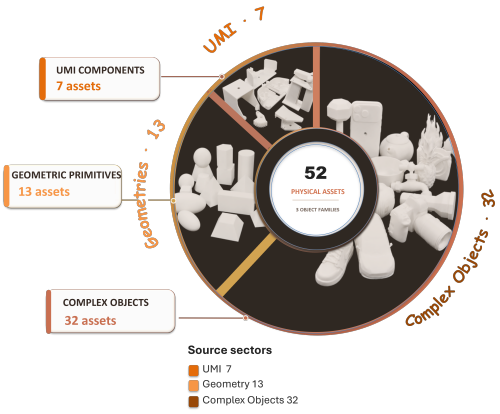

Figure 6: Object sources and category distribution in the real dataset, comprising 52 objects across 674 recordings.

For Daimon, we allocate 565 recordings to training, 18 to validation, and 91 to testing. All recordings of the seven test objects are excluded from training and validation, whereas validation uses separate recordings of training objects. For eFlesh, we use 60 recordings for training and hold out one physical instance from each of the six categories, excluding all recordings of these instances from training. This split evaluates generalization to unseen instances within known categories. Sensor-specific observation budgets and evaluation procedures are detailed in the corresponding experimental sections.

Evaluation Protocol and Metrics. We evaluate reconstruction using CD-L1 and F-score, and recognition using category accuracy or balanced accuracy. Evaluation is performed on held-out object instances as described above. The two simulation categories excluded from the training vocabulary are evaluated only for reconstruction. Because the datasets use different observation units—simulated grasps, Daimon observations, and eFlesh tactile episodes—we specify the observation budgets and sampling protocols in their corresponding experimental sections.

Few-touch performance in simulation. Figure 7 evaluates reconstruction and recognition on held-out instances from seen categories at $K = 1, 3, 5, 10$ grasps. We compare Elephant with PointSea, Simba, and SymmCompletion under matched observation budgets, adding the same classification head and category supervision to each baseline.

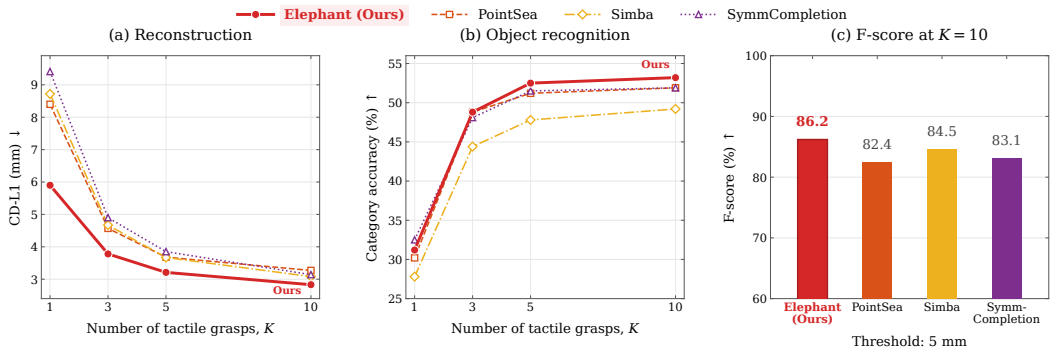

Figure 7: Simulation results on held-out instances from seen categories: CD-L1 and category accuracy across grasp budgets, and F-score at $K = 10$.

Elephant achieves the lowest CD-L1 across all evaluated budgets and the highest F-score at $K = 10$. Additional grasps improve both tasks, with recognition saturating earlier while reconstruction continues to benefit from increased contact coverage.

Real-world performance on Daimon. Table 2 compares reconstruction and recognition on held-out Daimon objects. Elephant achieves the lowest CD-L1 at three of four budgets, reaching 6.19 mm and 42.8% balanced accuracy at $K = 10$.

(a) Few-touch reconstruction

Method	Venue / Year	CD-L1 (mm) ↓				F-score (%) ↑
		$K = 1$	$K = 3$	$K = 5$	$K = 10$	$K = 10$
PCN (Yuan et al., 2018)	3DV'18	12.10	8.96	7.82	7.52	47.92
SnowflakeNet (Xiang et al., 2021)	ICCV'21	8.97	7.38	6.97	6.65	54.56
AdaPoinTr (Yu et al., 2023)	TPAMI'23	8.82	7.39	6.95	6.49	53.70
ProtoComp (Yu et al., 2024)	ECCV'24	31.09	24.19	23.31	22.42	5.20
SymmCompletion (Yan et al., 2025)	AAAI'25	12.62	9.47	8.51	7.87	46.15
ESCAPE (Bekci et al., 2025)	CVPR'25	25.19	18.13	15.00	13.71	22.47
PointSea (Zhu et al., 2025)	IJCV'25	12.46	9.30	8.62	8.12	45.74
Hyper-PCN (Li et al., 2026)	CVPR'26	9.42	**7.12**	6.95	6.48	53.53
Elephant (Ours)	–	**8.53**	7.32	**6.38**	**6.19**	**55.48**

(b) Few-touch object recognition

Method	Venue / Year	Balanced accuracy (%) ↑			
		$K = 1$	$K = 3$	$K = 5$	$K = 10$
PCN + Cls. (Yuan et al., 2018)	3DV'18	30.8	28.2	35.9	42.1
SnowflakeNet + Cls. (Xiang et al., 2021)	ICCV'21	7.7	17.9	20.5	23.1
AdaPoinTr + Cls. (Yu et al., 2023)	TPAMI'23	15.4	17.9	28.2	30.8
ProtoComp + Cls. (Yu et al., 2024)	ECCV'24	15.4	28.2	33.3	28.2
SymmCompletion + Cls. (Yan et al., 2025)	AAAI'25	7.7	30.8	33.3	33.3
ESCAPE + Cls. (Bekci et al., 2025)	CVPR'25	**33.3**	25.6	12.8	20.5
PointSea + Cls. (Zhu et al., 2025)	IJCV'25	7.7	15.4	17.9	15.4
Hyper-PCN + Cls. (Li et al., 2026)	CVPR'26	20.5	17.9	25.6	35.9
Elephant (Ours)	–	**33.3**	**38.5**	**41.0**	**42.8**

Table 2: Few-touch reconstruction and recognition on the real-world Daimon dataset. (a) Shape reconstruction evaluated by CD-L1 and F-score. (b) Semantic category recognition evaluated by balanced accuracy. K denotes the number of tactile observations. Best results are highlighted in bold.

4.2 Cross Sensors' Performance

We further evaluate Elephant using eFlesh, a magnetic tactile sensor with a measurement space substantially different from the optical Daimon sensor. We adapt the sensor-specific input encoder to eFlesh measurements while retaining the same object-level aggregation, reconstruction, and recognition framework. This setting evaluates whether Elephant can support few-touch object inference across different tactile sensing technologies.

Experimental setup. The eFlesh dataset contains 18 object instances from six semantic categories. We hold out one complete instance from each category and exclude all of its recordings from training. During training, each input is constructed from a single recording by sampling between 5 and 20 tactile episodes, with one valid time step selected per episode and observations from both fingers retained.

Observation budgets and evaluation. While each Daimon contact provides multiple local surface points, each eFlesh sensor provides a single 15-dimensional magnetic vector per sampled contact. We therefore use larger observation budgets of $K = 5, 10, 15, 20$ episodes to study how accumulating these sparse measurements supports object reconstruction and recognition. For each held-out instance, we construct three randomized sampling groups from one recording containing at least 20 valid episodes. Within each group, the four observation budgets are nested prefixes of the same randomly ordered episode sequence, enabling a controlled comparison as additional tactile evidence is introduced. This yields 18 sampled inputs at each observation budget across the six held-out instances.

(a) Few-touch reconstruction

Method	Venue / Year	CD-L1 (mm) ↓				F-score (%) ↑
		$K = 5$	$K = 10$	$K = 15$	$K = 20$	$K = 20$
PCN (Yuan et al., 2018)	3DV'18	7.50	7.19	6.96	7.22	46.96
SnowflakeNet (Xiang et al., 2021)	ICCV'21	**6.96**	**6.10**	**6.22**	6.33	51.82
AdaPoinTr (Yu et al., 2023)	TPAMI'23	8.18	6.55	6.32	6.75	49.65
ProtoComp (Yu et al., 2024)	ECCV'24	10.90	8.86	8.90	9.82	33.82
SymmCompletion (Yan et al., 2025)	AAAI'25	10.25	8.43	7.69	7.56	29.50
ESCAPE (Bekci et al., 2025)	CVPR'25	10.88	10.20	9.73	9.67	24.00
PointSea (Zhu et al., 2025)	IJCV'25	8.04	6.98	6.90	7.26	43.51
Hyper-PCN (Li et al., 2026)	CVPR'26	11.33	8.58	7.84	7.80	28.81
Elephant (Ours)	–	7.61	6.24	6.53	**6.22**	**56.21**

(b) Few-touch object recognition

Method	Venue / Year	Category accuracy (%) ↑				Balanced acc. (%) ↑
		$K = 5$	$K = 10$	$K = 15$	$K = 20$	$K = 20$
PCN + Cls. (Yuan et al., 2018)	3DV'18	55.56	61.11	61.11	61.11	61.11
SnowflakeNet + Cls. (Xiang et al., 2021)	ICCV'21	44.44	55.56	55.56	61.11	61.11
AdaPoinTr + Cls. (Yu et al., 2023)	TPAMI'23	61.11	61.11	**77.78**	66.67	66.67
ProtoComp + Cls. (Yu et al., 2024)	ECCV'24	**66.67**	61.11	55.56	50.00	50.00
SymmCompletion + Cls. (Yan et al., 2025)	AAAI'25	61.11	44.44	50.00	50.00	50.00
ESCAPE + Cls. (Bekci et al., 2025)	CVPR'25	33.33	38.89	33.33	38.89	38.89
PointSea + Cls. (Zhu et al., 2025)	IJCV'25	50.00	55.56	61.11	50.00	50.00
Hyper-PCN + Cls. (Li et al., 2026)	CVPR'26	38.89	44.44	38.89	33.33	33.33
Elephant (Ours)	–	50.00	**77.78**	**77.78**	**94.44**	**94.44**

Table 3: Few-touch reconstruction and recognition on the real-world eFlesh dataset. We evaluate six held-out instances, one per category, with three sampled subsets per instance, yielding 18 evaluations at each budget $K \in \{5, 10, 15, 20\}$. Each episode contributes paired left/right observations at one timestamp. (a) Reconstruction is evaluated by CD-L1 (mm) and F-score at a 5 mm threshold. (b) Category accuracy equals balanced accuracy because category sample counts are equal. Best results are highlighted in bold, including ties.

Comparison with Daimon. Elephant supports joint reconstruction and recognition with both Daimon's local surface measurements and eFlesh's magnetic observations. Its reconstruction advantage emerges at different budgets: Elephant achieves the lowest CD-L1 at three of the four evaluated Daimon budgets, whereas on eFlesh it outperforms the reconstruction baselines only at $K = 20$.

Both evaluations show diminishing reconstruction gains at larger budgets, but recognition follows different trends. On Daimon, increasing K from 5 to 10 reduces CD-L1 from 6.38 to 6.19 mm and raises balanced accuracy from 41.0% to 42.8%. On eFlesh, increasing K from 10 to 20 yields little net change in CD-L1, from 6.24 to 6.22 mm, while recognition accuracy rises from 77.78% to 94.44%. These results demonstrate that the same object-level inference architecture, paired with an eFlesh-specific encoder, supports both shape reconstruction and category recognition from sparse magnetic tactile measurements.

5 Conclusion

We presented Elephant, a framework for joint object reconstruction and recognition from a small number of spatially localized tactile observations. By aligning contacts in a common object frame and combining geometric evidence with semantic reasoning, Elephant infers both complete shape and object category without visual input at inference time. Experiments in simulation and on real Daimon and eFlesh sensors show that a few contacts already provide substantial object-level information, while additional observations further reduce geometric ambiguity. The results also demonstrate that the same object-level inference framework can operate across distinct tactile sensing technologies through sensor-specific encoders. Overall, our study highlights the potential of sparse touch as a standalone source for global object understanding.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing, grammatical correction, LaTeX editing, and limited code debugging. They were not used to generate experimental results or make scientific conclusions. All research ideas, methodological design, experiments, analyses, and reported results were developed, conducted, and verified by the authors. The authors take full responsibility for the content of the paper.

REPRODUCIBILITY STATEMENT

We provide the architectural, training, dataset, and evaluation details required to reproduce the reported experiments. The supplementary material includes preprocessing and calibration procedures, implementation hyperparameters, data splits, observation-sampling protocols, baseline configurations, and metric definitions. All evaluations use fixed splits and held-out object instances as described in the paper. We plan to release the training and evaluation code, configuration and split files, model checkpoints, and releasable processed tactile data and calibration metadata to facilitate reproduction of the simulation and real-world experiments.

REFERENCES

Iretiayo Akinola, Jie Xu, Jan Carius, Dieter Fox, and Yashraj Narang. Tacsl: A library for visuotactile sensor simulation and learning. *IEEE Transactions on Robotics*, 41:2645–2661, 2025. doi: 10.1109/TRO.2025.3547267.

Osher Azulay, Alon Mizrahi, Nimrod Curtis, and Avishai Sintov. Augmenting tactile simulators with real-like and zero-shot capabilities. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5702–5708, 2024. doi: 10.1109/ICRA57147.2024.10610442.

Burak Bekci, Nassir Navab, Federico Tombari, and Mahdi Saleh. Escape: Equivariant shape completion via anchor point encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6480–6489, June 2025.

Raunaq Bhirangi, Venkatesh Pattabiraman, Enes Erciyes, Yifeng Cao, Tess Hellebrekers, and Lerrel Pinto. Anyskin: Plug-and-play skin sensing for robotic touch. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 16563–16570, 2025. doi: 10.1109/ICRA55743.2025.11128638.

Berk Calli, Arjun Singh, Aaron Walsman, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M. Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. In *2015 International Conference on Advanced Robotics (ICAR)*, pp. 510–517, 2015. doi: 10.1109/ICAR.2015.7251504.

Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012, 2015. URL https://arxiv.org/abs/1512.03012.

Weihang Chen, Yuan Xu, Zhenyang Chen, Peiyu Zeng, Renjun Dang, Rui Chen, and Jing Xu. Bidirectional sim-to-real transfer for GelSight tactile sensors with CycleGAN. *IEEE Robotics and Automation Letters*, 7(3):6187–6194, 2022. doi: 10.1109/LRA.2022.3167064.

Weihang Chen, Jing Xu, Fanbo Xiang, Xiaodi Yuan, Hao Su, and Rui Chen. General-purpose sim2real protocol for learning contact-rich manipulation with marker-based visuotactile sensors. *IEEE Transactions on Robotics*, 40:1509–1526, 2024. doi: 10.1109/TRO.2024.3352969.

Alex Church, John Lloyd, Raia Hadsell, and Nathan F. Lepora. Tactile sim-to-real policy transfer via real-to-sim image translation. In Aleksandra Faust, David Hsu, and Gerhard Neumann (eds.), *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pp. 1645–1654. PMLR, 08–11 Nov 2022. URL https://proceedings.mlr.press/v164/church22a.html.

Mauro Comi, Yijiong Lin, Alex Church, Alessio Tonioni, Laurence Aitchison, and Nathan F. Lepora. TouchSDF: A DeepSDF approach for 3D shape reconstruction using vision-based tactile sensing. *IEEE Robotics and Automation Letters*, 9(6):5719–5726, 2024. doi: 10.1109/LRA.2024.3396054.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13142–13153, June 2023.

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 *2022 International Conference on Robotics and Automation (ICRA)*, pp. 2553–2560, 2022. doi: 10.1109/ICRA46639.2022.9811809.

Irving Fang, Kairui Shi, Xujin He, Siqi Tan, Yifan Wang, Hanwen Zhao, Hung-Jui Huang, Wenzhen Yuan, Chen Feng, and Jing Zhang. Fusionsense: Bridging common sense, vision, and touch for robust sparse-view reconstruction. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 15798–15805, 2025. doi: 10.1109/ICRA55743.2025.11128188.

Ruoxuan Feng, Jiangyu Hu, Wenke Xia, Tianci Gao, Ao Shen, Yuhao Sun, Bin Fang, and Di Hu. Anytouch: Learning unified static-dynamic representation across multiple visuo-tactile sensors. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 31265–31285, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/4d893f766ab60e5337659b9e71883af4-Paper-Conference.pdf.

Langzhe Gu, Hung-Jui Huang, Mohamad Qadri, Michael Kaess, and Wenzhen Yuan. Touchanything: Diffusion-guided 3d reconstruction from sparse robot touches. In *European Conference on Computer Vision*, 2026. URL https://eccv.ecva.net/virtual/2026/poster/4442.

Harsh Gupta, Yuchen Mo, Shengmiao Jin, and Wenzhen Yuan. Sensor-invariant tactile representation. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 96515–96538, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/f05a11e682bf9c57ab62f7cff08fbe5b-Paper-Conference.pdf.

Jian Hou, Xin Zhou, Qihan Yang, and Adam J. Spiers. UniTac-NV: A unified tactile representation for non-vision-based tactile sensors. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 17854–17860, 2025. doi: 10.1109/IROS60139.2025.11247617.

Tudor Jianu, Daniel Fernandes Gomes, and Shan Luo. Reducing tactile sim2real domain gaps via deep texture generation networks. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 8305–8311, 2022. doi: 10.1109/ICRA46639.2022.9811801.

Xingshuo Jing and Kun Qian. Reducing cross-sensor domain gaps in tactile sensing via few-sample-driven style-to-content unsupervised domain adaptation. *Sensors*, 25(1):256, 2025. ISSN 1424-8220. doi: 10.3390/s25010256. URL https://www.mdpi.com/1424-8220/25/1/256.

Arunim Joarder, Arjun Bhardwaj, René Zurbrügg, Mayank Mittal, Florin Püntener, Sira Bielefeldt, Cosmin Roman, Vaishakh Patil, and Marco Hutter. Tactspace: Learning a physics-enriched shared latent space for tactile sim-to-real transfer. arXiv preprint arXiv:2606.18959, 2026. URL https://arxiv.org/abs/2606.18959.

Linfei Li, Pei Tan, Siqi Li, Changqing Zou, and Yue Gao. Hyper-pcn: Hypergraph-based point cloud completion via high-order correlation modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 39121–39130, June 2026.

Yangrong Liu, Jian Li, Huaiyu Wang, Ming Lu, Haorao Shen, and Qin Wang. TAPCNet: Tactile-assisted point cloud completion network via iterative fusion strategy. *IET Computer Vision*, 19 (1):e70012, 2025. doi: 10.1049/cvi2.70012. URL https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/cvi2.70012.

Prajval Kumar Murali, Bernd Porr, and Mohsen Kaboli. Touch if it's transparent! actor: Active tactile-based category-level transparent object reconstruction. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10792–10799, 2023. doi: 10.1109/IROS55552.2023.10341680.

Sameer Pai, Tao Chen, Megha Tippur, Edward Adelson, Abhishek Gupta, and Pulkit Agrawal. Tactofind: A tactile only system for object retrieval. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8025–8032, 2023. doi: 10.1109/ICRA48891.2023.10160289.

Junsung Park, Sachin Bhadang, Carmelo Sferrazza, Sha Yi, and Xiaolong Wang. Tactx: Learning shared tactile representations across diverse sensors. arXiv preprint arXiv:2606.31236, 2026. URL https://arxiv.org/abs/2606.31236.

Venkatesh Pattabiraman, Zizhou Huang, Daniele Panozzo, Denis Zorin, Lerrel Pinto, and Raunaq Bhirangi. eFlesh: Highly customizable magnetic touch sensing using cut-cell microstructures. arXiv preprint arXiv:2506.09994, 2025. URL https://arxiv.org/abs/2506.09994.

Samanta Rodriguez, Yiming Dou, Miquel Oller, Andrew Owens, and Nima Fazeli. Cross-sensor touch generation. In Joseph Lim, Shuran Song, and Hae-Won Park (eds.), *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 152–167. PMLR, 27–30 Sep 2025. URL https://proceedings.mlr.press/v305/rodriguez25a.html.

Lukas Rustler, Jens Lundell, Jan Kristof Behrens, Ville Kyrki, and Matej Hoffmann. Active visuo-haptic object shape completion. *IEEE Robotics and Automation Letters*, 7(2):5254–5261, 2022. doi: 10.1109/LRA.2022.3152975.

Carmelo Sferrazza, Thomas Bi, and Raffaello D'Andrea. Learning the sense of touch in simulation: a sim-to-real strategy for vision-based tactile sensing. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4389–4396, 2020. doi: 10.1109/IROS45743.2020.9341285.

Zilin Si and Wenzhen Yuan. Taxim: An example-based simulation model for GelSight tactile sensors. *IEEE Robotics and Automation Letters*, 7(2):2361–2368, 2022. doi: 10.1109/LRA.2022.3142412.

Edward Smith, Roberto Calandra, Adriana Romero, Georgia Gkioxari, David Meger, Jitendra Malik, and Michal Drozdzal. 3d shape reconstruction from vision and touch. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 14193–14206. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/a3842ed7b3d0fe3ac263bcabd2999790-Paper.pdf.

Edward Smith, David Meger, Luis Pineda, Roberto Calandra, Jitendra Malik, Adriana Romero Soriano, and Michal Drozdzal. Active 3d shape reconstruction from vision and touch. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 16064–16078. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/8635b5fd6bc675033fb72e8a3ccc10a0-Paper.pdf.

S. Suresh, Z. Si, J. Mangelson, W. Yuan, and M. Kaess. ShapeMap 3-D: Efficient shape mapping through dense touch and vision. In *Proc. IEEE Intl. Conf. on Robotics and Automation, ICRA*, pp. 7073–7080, Philadelphia, PA, USA, May 2022.

Sudharshan Suresh, Haozhi Qi, Tingfan Wu, Taosha Fan, Luis Pineda, Mike Lambeta, Jitendra Malik, Mrinal Kalakrishnan, Roberto Calandra, Michael Kaess, Joseph Ortiz, and Mustafa Mukadam. Neuralfeels with neural fields: Visuotactile perception for in-hand manipulation. *Science Robotics*, 9(96):eadl0628, 2024. doi: 10.1126/scirobotics.adl0628. URL https://www.science.org/doi/abs/10.1126/scirobotics.adl0628.

Yuanbo Wang, Zhaoxuan Zhang, Jiajin Qiu, Dilong Sun, Zhengyu Meng, Xiaopeng Wei, and Xin Yang. Touch2shape: Touch-conditioned 3d diffusion for shape exploration and reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5656–5665, June 2025.

Qiwei Wu, Xuanbin Peng, Jiayu Zhou, Zhuoran Sun, Xiaogang Xiong, and Yunjiang Lou. Rttf: Rapid tactile transfer framework for contact-rich manipulation tasks. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2913–2920, 2024. doi: 10.1109/IROS58592.2024.10801764.

Peng Xiang, Xin Wen, Yu-Shen Liu, Yan-Pei Cao, Pengfei Wan, Wen Zheng, and Zhizhong Han. SnowflakeNet: Point cloud completion by snowflake point deconvolution with Skip-Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5499–5509, October 2021.

Jingxi Xu, Shuran Song, and Matei Ciocarlie. Tandem: Learning joint exploration and decision making with tactile sensors. *IEEE Robotics and Automation Letters*, 7(4):10391–10398, 2022. doi: 10.1109/LRA.2022.3193466.

Hongyu Yan, Zijun Li, Kunming Luo, Li Lu, and Ping Tan. Symmcompletion: High-fidelity and high-consistency point cloud completion with symmetry guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2025. doi: 10.1609/aaai.v39i9.32984. URL https://ojs.aaai.org/index.php/AAAI/article/view/32984.

Fengyu Yang, Chao Feng, Ziyang Chen, Hyoungseob Park, Daniel Wang, Yiming Dou, Ziyao Zeng, Xien Chen, Rit Gangopadhyay, Andrew Owens, and Alex Wong. Binding touch to everything: Learning unified multimodal tactile representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26340–26353, June 2024.

Xumin Yu, Yongming Rao, Ziyi Wang, Jiwen Lu, and Jie Zhou. AdaPoinTr: Diverse point cloud completion with adaptive geometry-aware transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):14114–14130, 2023. doi: 10.1109/TPAMI.2023.3309253.

Xumin Yu, Yanbo Wang, Jie Zhou, and Jiwen Lu. Protocomp: Diverse point cloud completion with controllable prototype. In *European Conference on Computer Vision*. Springer, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/6685_ECCV_2024_paper.php.

Wentao Yuan, Tejas Khot, David Held, Christoph Mertz, and Martial Hebert. PCN: Point completion network. In *2018 International Conference on 3D Vision (3DV)*, pp. 728–737, 2018. doi: 10.1109/3DV.2018.00088.

Mabel M. Zhang, Nikolay Atanasov, and Kostas Daniilidis. Active end-effector pose selection for tactile object recognition through monte carlo tree search. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3258–3265, 2017. doi: 10.1109/IROS.2017.8206161.

Jialiang Zhao, Yuxiang Ma, Lirui Wang, and Edward Adelson. Transferable tactile transformers for representation learning across diverse sensors and tasks. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard (eds.), *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 3766–3779. PMLR, 06–09 Nov 2025. URL https://proceedings.mlr.press/v270/zhao25c.html.

Yongqiang Zhao, Kun Qian, Boyi Duan, and Shan Luo. Fots: A fast optical tactile simulator for sim2real learning of tactile-motor robot manipulation skills. *IEEE Robotics and Automation Letters*, 9(6):5647–5654, 2024. doi: 10.1109/LRA.2024.3396665.

Haodong Zheng, Andrei Jalba, Raymond H. Cuijpers, Wijnand IJsselsteijn, and Sanne Schoenmakers. A bayesian framework for active tactile object recognition, pose estimation and shape transfer learning. arXiv preprint arXiv:2409.06912, 2026. URL https://arxiv.org/abs/2409.06912.

Shaohong Zhong, Alessandro Albini, Perla Maiolino, and Ingmar Posner. Tactgen: Tactile sensory data generation via zero-shot sim-to-real transfer. *IEEE Transactions on Robotics*, 41:1316–1328, 2025. doi: 10.1109/TRO.2024.3521967.

Zhe Zhu, Honghua Chen, Xing He, and Mingqiang Wei. Pointsea: Point cloud completion via self-structure augmentation. *International Journal of Computer Vision*, 133(7):4770–4794, 2025. doi: 10.1007/s11263-025-02400-y. URL https://doi.org/10.1007/s11263-025-02400-y.