EGOFACT: ESTIMATING DENSE FULL-HAND TACTILE FIELDS FROM EGOCENTRIC HUMAN VIDEOS

Anonymous authors
Paper under double-blind review

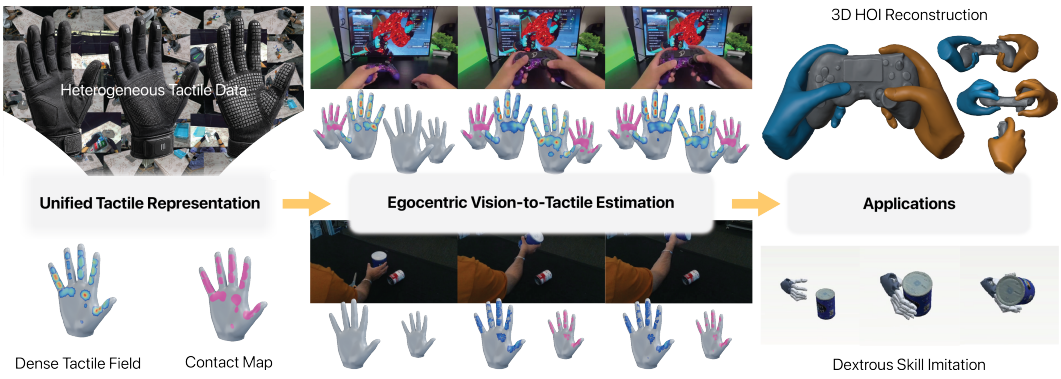

Figure 1: **We present EgoFact:** a unified framework that *(i.)* represents heterogeneous tactile data in a unified 3D dense filed over hand; then *(ii.)* estimates tactile fields from egocentric vision; and *(iii.)* applies it to downstream 3D HOI reconstruction and dextrous skill imitation in simulation.

ABSTRACT

Understanding human dexterity from egocentric videos requires the ability to recover not only the kinematics, but also the physical tactile signals underlying the hand-object interface. However, estimating such tactile signals from egocentric videos remains a challenging task due to severe occlusions and heterogeneous tactile representations across sensing devices and datasets. To bridge this gap, we introduce **EgoFact**, a unified framework for estimating dense, full-hand tactile fields of both hands directly from egocentric videos. Specifically, **1)** we map heterogeneous tactile measurements onto a canonical 3D hand surface with spatial decay and propagation around sensor centers, yielding a unified full-hand tactile representation for cross-device training and evaluation; **2)** we develop a tracker-free vision-to-tactile model with force-guided learning objective to effectively infer tactile fields directly from RGB inputs; and **3)** we introduce EgoFact-Data, a cross-device egocentric-tactile dataset with 100 hours of recordings, 800 clips, three glove types, six scenes, and 200 manipulation tasks, together with scene- and task-held-out protocols for evaluating generalization. Extensive experiments across five egocentric–tactile benchmarks demonstrate consistent improvements over existing baselines, achieving up to around 25% higher contact accuracy, 30% higher field IoU, and 40% lower force error. Notably, EgoFact further improves 3D hand–object interaction reconstruction and supports dexterous skill imitation, presenting the broader value of vision-to-tactile estimation for physically grounded egocentric intelligence.

1 INTRODUCTION

The tactile states underlying observable actions are critical to understanding human dexterity from a first-person perspective (Yamaguchi & Atkeson, 2019; Suomalainen et al., 2022), as they characterize where contact occurs, how force is distributed, and how these patterns evolve over time. Recovering tactile signals from egocentric vision thus plays a crucial role in connecting visual observations to physical interactions, supporting physically grounded action understanding, interaction reconstruction, and learning from human dexterity.

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.

However, training and evaluating vision-to-tactile models across datasets in a unified protocal is challenging. Although recent egocentric-tactile datasets (Song et al., 2025; Zhou et al., 2026; Zeng et al., 2026; Lyu et al., 2026; Cao et al., 2026) have substantially advanced vision-to-tactile learning, their sensors differ in placement, spatial coverage, and response characteristics, so their measurements are organized in heterogeneous formats. These discrepancies constrain prior methods to training and evaluation within individual datasets, without a unified protocol across datasets. Concurrent works (Jeon et al., 2026; Zhang et al., 2026) address this limitation by assigning sensor readings to a canonical MANO mesh (Romero et al., 2022) and directly averaging values in overlapping regions. However, this scheme remains device-dependent and produces piecewise-constant patches without modeling spatial force decay and propagation around sensor locations. EgoPressure (Zhao et al., 2025) and EgoTactile (Zeng et al., 2026) instead represent tactile signals in UV and pixel space respectively, without explicitly accounting for the geometric connectivity of the 3D hand surface.

Another challenge of vision-to-tactile estimation lies in the ill-posed nature of inferring unobserved tactile signals from visual context. A natural approach is to first recover hand kinematics from vision and use them to guide tactile estimation (Zhou et al., 2026; Jeon et al., 2026). However, hand kinematics do not capture the interaction interface and are ambiguous for informing the underlying tactile signals. Requiring kinematics as input also propagates errors from upstream hand trackers into tactile estimation. Concurrent work (Zhang et al., 2026) explores a tracker-free method, but lacks dedicated architectural design and systematic evaluation. How best to learn tactile features directly from RGB inputs without upstream hand trackers therefore remains an open question.

To address these challenges, we introduce EgoFact, a framework for estimating dense tactile fields over the 3D hand surface of both hands from egocentric videos. Its first component is a unified tactile representation for heterogeneous tactile data across datasets and sensing devices. Using pre-defined correspondences between sensors and the canonical MANO mesh (Romero et al., 2022), we center a geodesic Gaussian kernel at each sensor location and aggregate kernels by taking the upper envelope. Geodesic distance models spatial force decay along the hand surface, and the upper-envelope aggregation propagates sensor measurements to full-hand tactile fields by retaining the largest sensor response at each location. The resulting fields share the same geometry domain and output dimensionality across any tactile datasets and devices. Building on this representation, we develop a vision-to-tactile model that effectively fuses multi-level visual features with learned tactile features, and estimates tactile fields directly from RGB video without upstream hand trackers. This avoids inheriting errors from a separate pose-recovery stage, including when gloves or occlusion hinder detection and tracking. During training, instead of using uniformly weighted loss, we introduce a force-guided learning objective to prioritize supervisions within statistically populated intervals, thereby promoting more stable training and reducing sensitivity to noise at the extremes.

To diversify tactile supervision and broaden evaluation, we collect EgoFact-Data, a cross-device egocentric-tactile dataset comprising 100 hours of recordings across 800 clips, acquired with three glove types in six scenes and covering 200 manipulation tasks. Raw measurements from all gloves are converted into our unified representation. Complementing existing evaluation protocols that focus on subjects and interacting objects (Zhou et al., 2026; Zeng et al., 2026), we establish scene-held-out and task-held-out protocols to assess transfer to recording environments and manipulation tasks excluded from training, respectively.

Extensive experiments on OpenTouch, EgoTouch, EgoTactile, HumanTouch, and EgoFact-Data demonstrate that EgoFact significantly improves contact estimation accuracy, field estimation IoU, and absolute force error relative to baselines. Ablation studies systematically validate our three key designs: *(i.)* our representation models spatial force decay and propagation while preserving smoothness and peaks; *(ii.)* the tracker-free model and tactile feature learning strategy achieve better performance among the alternatives; and *(iii.)* the force-guided learning objective outperforms uniformly weighted losses. We further apply the trained model to in-the-wild cases, showing its applicability to unconstrained inputs and its utility for various downstream applications, such as refining 3D hand-object interaction reconstruction and guiding dexterous skill imitation learning.

To summarize, our contributions are threefold:

- We introduce a unified tactile representation using geodesic Gaussian kernels and upper-envelope aggregation on a canonical 3D hand surface, unifying heterogeneous tactile data into a shared domain with spatial force decay and better response peak preserving.

- We develop a hand tracker-free vision-to-tactile model together with a force-guided training objective, enabling accurate and dense full-hand tactile estimation without requirement of hand trackers.
- We collect EgoFact-Data, a cross-device egocentric-tactile dataset spanning 100 hours, 800 clips, three glove types, six scenes, and 200 manipulation tasks, and establish protocols for evaluating transfer to unseen scenes and tasks. Experiments on four existing benchmarks and EgoFact-Data, together with applications on in-the-wild videos, assess our framework's effectiveness and generalization.

2 Related Works

Full-hand Tactile Representation. Existing full-hand tactile representations range from sensor grids and canonical heatmaps (Song et al., 2025; Zhou et al., 2026; Zeng et al., 2026) to UV maps and vertex-wise fields on hand meshes (Zhao et al., 2025; Jeon et al., 2026; Zhang et al., 2026). These hand-centric formulations facilitate learning from heterogeneous annotations, with spatial mappings based on weighted interpolation or averaging followed by geodesic smoothing (Jeon et al., 2026; Zhang et al., 2026). We combine geodesic Gaussian kernels with upper-envelope aggregation on a canonical 3D hand surface, providing a shared tactile domain that models spatial force decay while preserving strong local responses.

Full-Hand Tactile Estimation. Early approaches infer hand or fingertip pressure from RGB images under planar-contact or restricted interaction settings (Grady et al., 2022; 2024; Zhao et al., 2025). Recent methods extend this task to full-hand tactile estimation in free object manipulation (Zhou et al., 2026; Zeng et al., 2026; Jeon et al., 2026; Zhang et al., 2026; Ilic et al., 2026). These approaches rely on either hand pose priors or auxiliary synthetic supervision, trained and evaluated only on individual datasets. We develop a hand tracker-free vision-to-tactile model that are trained purely on tactile data, and can be evaluated across datasets.

Egocentric Full-Hand Tactile Dataset. Recent egocentric-tactile datasets extend contact and tactile supervision from geometry-based synthetic data (Taheri et al., 2020; Brahmbhatt et al., 2019; Fan et al., 2023; Yang et al., 2022; Chao et al., 2021b; Liu et al., 2022; Zhao et al., 2024) and planar interactions (Zhao et al., 2025; Grady et al., 2024) to full-hand manipulation across daily environments (Song et al., 2025; Zhou et al., 2026; Zeng et al., 2026; Lyu et al., 2026). Although great efforts have been made, there still lacks scalability and diversity of supervision and evaluation. We complement this limitation by collecting a cross-device egocentric-tactile dataset with evaluation protocols for tactile learning across sensing systems and generalization to new interaction contexts.

3 Method

Our goal is to estimate dense full-hand tactile fields for both hands from egocentric videos. We first introduce a unified tactile representation that maps heterogeneous tactile data onto a canonical 3D hand surface using geodesic Gaussian kernels and upper-envelope aggregation (Sec. 3.1). Building on this representation, we develop a vision-to-tactile model that estimates dense full-hand tactile fields directly from RGB inputs, without requiring upstream hand trackers (Sec. 3.2). To increase the diversity of tactile supervisions and broaden evaluations on this task, we further collect a cross-device egocentric-tactile dataset using three different tactile gloves (Sec. 3.3).

3.1 Unified Full-Hand Tactile Representation

Canonical Tactile Template. We follow OpenTouch (Song et al., 2025) to use a subdivided right-hand MANO model (Romero et al., 2022) in a fixed, flat-hand pose to define a canonical geometry domain. Let $\boldsymbol{V}$ denote the vertex set of the subdivided hand mesh, as illustrated in Fig. 2, we manually select the vertices on the palmar side to form the canonical tactile template $\boldsymbol{V}_{\text{tac}} \subset \boldsymbol{V}$. The corresponding mesh edges $\boldsymbol{E}_{\text{tac}}$ are retained to describe connectivity within $\boldsymbol{V}_{\text{tac}}$. The same template serves as the shared geometry domain for both hands.

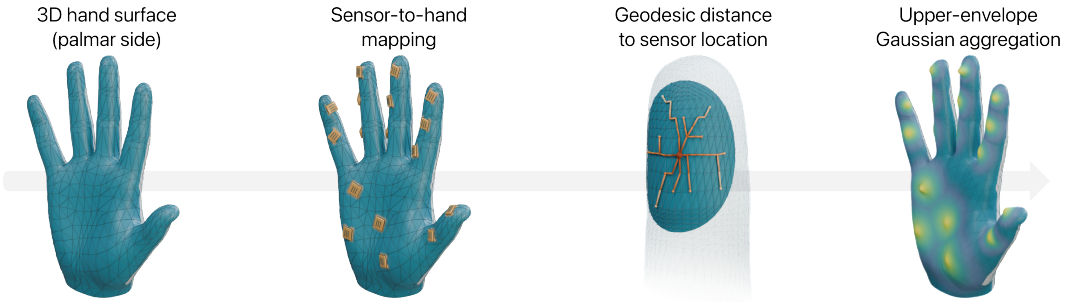

Figure 2: **Representation Overview.** Starting from *(i.)* the canonical 3D hand surface, *(ii.)* we build correspondences between sensors and palmar regions, then *(iii.)* define geodesic Gaussian kernels at sensor centers, and finally *(iv.)* convert sensor measurements to dense full-hand tactile fields.

Tactile Sensor-to-Template Mapping. To establish correspondences between tactile sensors and the template $\boldsymbol{V}_{\text{tac}}$, we manually define a fixed tactile sensor-to-template mapping for each device type. Formally, given device with K sensors in any placement, we define $\mathcal{M}(s_k) = \boldsymbol{V}_{s_k} \subset \boldsymbol{V}_{\text{tac}}$ for each sensor, where $\boldsymbol{V}_{s_k}$ is the nonempty set of template vertex indices corresponding to the region covered by the k-th sensor s_k. Separate mappings are constructed for the left and right gloves, which are aligned to the same regions on $\boldsymbol{V}_{\text{tac}}$. Meanwhile, We define the center of s_k as $\boldsymbol{c}_k = \frac{1}{|\boldsymbol{V}_{s_k}|} \sum_{\boldsymbol{v}_i \in \boldsymbol{V}_{s_k}} \boldsymbol{v}_i$, which is the centroid of its covering region $\boldsymbol{V}_{s_k}$.

Dense Tactile Field Based on Geodesic Gaussian. We then construct dense tactile fields from discrete tactile sensors by defining multiple geodesic Gaussian kernels on the template and aggregating them via the upper envelope. Existing methods (Song et al., 2025; Jeon et al., 2026; Zhang et al., 2026) adopt a naive assignment scheme by copying each sensor reading to all vertices in its corresponding region and averaging readings where regions overlap. However, this is device-dependent and produces piecewise-constant patches. EgoTactile (Zeng et al., 2026) utilizes row-normalized Gaussian interpolation in pixel space on a 2D hand canvas, ignoring surface connectivity of hands in 3D. Our representation is aimed to simulate the force distribution on 3D hand surface, modeling the spatial decay around sensor centers and propagate to locations that are not covered by sensors.

Specifically, we first build an undirected graph $\boldsymbol{G} = (\boldsymbol{V}_{\text{tac}}, \boldsymbol{E}_{\text{tac}})$ based on our canonical tactile template. Let $\boldsymbol{D}_{\boldsymbol{G}}(\boldsymbol{v}_i, \boldsymbol{v}_j)$ denote the shortest-path distance between any two vertices $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ on $\boldsymbol{G}$, providing a discrete approximation to the surface geodesic distance. Then, since the sensor center $\boldsymbol{c}_k$ may not belong to $\boldsymbol{G}$, for each $\boldsymbol{v}_i \in \boldsymbol{V}_{\text{tac}}$, we calculate its distance to sensor s_k by:

$$\boldsymbol{D}(s_k, \boldsymbol{v}_i) = \min_{\boldsymbol{v}_j \in \boldsymbol{V}_{s_k}} \left[\|\mathbf{c}_k - \mathbf{v}_j\|_2 + \boldsymbol{D}_{\boldsymbol{G}}(\boldsymbol{v}_i, \boldsymbol{v}_j) \right]. \tag{1}$$

The minimization over $\boldsymbol{V}_{s_k}$ selects the shortest path. Subsequently, the Gaussian weight centered on $\boldsymbol{c}_k$ at each vertex $\boldsymbol{v}_i$ is computed according to $\boldsymbol{D}(s_k, \boldsymbol{v}_i)$:

$$\boldsymbol{W}(s_k, \boldsymbol{v}_i) = \exp\left(-\frac{\boldsymbol{D}(s_k, \boldsymbol{v}_i)^2}{2\sigma^2} \right), \tag{2}$$

where $\sigma = 5\text{mm}$. Assume that r_k is the min-max normalized reading of sensor s_k, we finally aggregate the k Gaussians and take the upper envelope as the resulting dense tactile field $\boldsymbol{y} \in \mathbb{R}^{|\boldsymbol{V}_{\text{tac}}|}$:

$$\boldsymbol{y}_i = \max_{k=1,\ldots,K} \left[\boldsymbol{W}(s_k, \boldsymbol{v}_i) r_k \right]. \tag{3}$$

3.2 Model Architecture

Input Preprocessing. Given a set of egocentric images, we first preprocess them by localizing the target hand in each frame and cropping the corresponding region. Since conventional detectors (Redmon et al., 2016; Xu et al., 2022; Potamias et al., 2025) suffer from gloved hands and severe occlusions, we adopt general-purpose segmentation model SAM3 (Carion et al., 2026) to segment the target hand and derive its bounding box from the outer contour of the segmented mask. To retain the hand-object interface as fully as possible, the bounding box is enlarged and we crop the image using the longer side. The cropped patch are then resized to obtain the input image

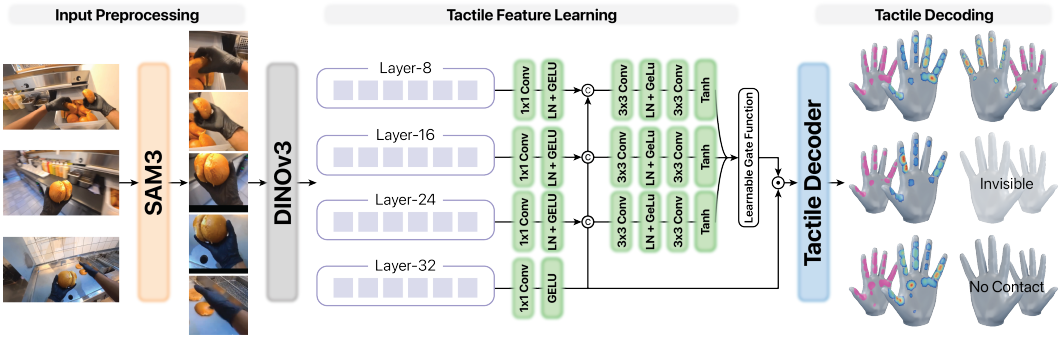

Figure 3: **Method Overview.** EgoFact *(i.)* takes egocentric RGB inputs, then *(ii.)* extracts and fuses multi-level features of general-purpose visual encoder to obtain tactile feature, and finally *(iii.)* decodes it to dense full-hand tactile fields.

$\boldsymbol{I} \in \mathbb{R}^{256\times192\times3}$. Note that images of left hand are horizontally flipped after cropping to ensure a consistent visual orientation (Pavlakos et al., 2024; Potamias et al., 2025; Zhang et al., 2025).

Tactile Feature Learning via Fusing Multi-Level Visual Features. We learn tactile feature from images by extracting and fusing multi-level visual features. Our key insight is, tactile signals is determined by frequently occluded hand-object interfaces, estimating them requires cues from low-level hand kinematics and object shapes, to high-level interaction contexts. Per-layer analyses (Siméoni et al., 2025) validate that features from different visual encoder layers provide complementary information, which are effective for dense prediction tasks when promptly fused (Ranftl et al., 2021).

Accordingly, as illustrated in Fig. 3, we differ from prior methods (Zhou et al., 2026; Jeon et al., 2026; Zhang et al., 2026) by retaining features $\boldsymbol{f}^{(l)} = \mathcal{F}(\boldsymbol{I}), l \in \{8, 16, 24, 32\}$ from multiple intermediate layers and the final layer of foundational DINOv3 encoder $\mathcal{F}$ (Siméoni et al., 2025). To learn informative contexts from each $\boldsymbol{f}^{(l)}$, we design a projection block consisting of a convolution module, a layer normalization (excluded from $\boldsymbol{f}^{(32)}$), and a GELU activation to obtain projected feature $\boldsymbol{p}^{(l)}$. Taking the projected final-layer feature $\boldsymbol{p}^{(32)}$ as the base feature, we concatenate intermediate ones with it along the channel dimension and learn the residual feature $\boldsymbol{r}^{(l)}, l \in \{8, 16, 24\}$ through a larger convolution kernel and an additional tanh activation. Subsequently, we update the base feature to obtain our target tactile feature $\boldsymbol{f}_{\text{tac}}$ by combining it with residuals through learnable weights and softmax gating:

$$\boldsymbol{f}_{\text{tac}} = \boldsymbol{p}^{(32)} + \gamma \sum_{l\in\{8,16,24\}} \alpha_l \boldsymbol{r}^{(l)}, \tag{4}$$

where α_l is the learnable gate factor of layer-l after softmax operation, and γ is the controlling factor for the overall magnitude of the residual updates.

Tactile Decoding And Force-Guided Training Objective. Given the learned tactile feature $\boldsymbol{f}_{\text{tac}}$, we flatten its spatial patch grid and feed it into a four-layer MLP decoder. A final sigmoid activation maps the decoder logits to normalized tactile estimation results $\hat{\boldsymbol{y}} \in \mathbb{R}^{|\boldsymbol{V}_{\text{tac}}|}$. During training, we observe that the learning process is unstable and sensitive to noisy tactile supervisions at the extremes when using uniformly weighted loss. Therefore, we define a piecewise-linear weighting profile $h(\cdot)$ which partitions the whole normalized value range to several intervals including $[0, 0.05), [0.05, 0.1), [0.1, 0.3), [0.3, 1]$, and assigns different weight to more populated intervals counted on the training corpus. In addition, a non-contact regularization term is introduced to enforce the model to estimate zeros when the target value is below a threshold ϵ, which may arise from calibration errors or device sensitivity. Consequently, our force-guided training objective is formulated as:

$$\mathcal{L} = \frac{1}{|\boldsymbol{V}_{\text{tac}}|} \sum_{i=0}^{|\boldsymbol{V}_{\text{tac}}|-1} \left(h\left(\boldsymbol{y}_i\right) \left\|\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i\right\|_2^2 + \max\left(\hat{\boldsymbol{y}}_i - \epsilon, 0\right)^2 \right). \tag{5}$$

For more details of our loss weighting scheme, please refer to Appendix.

Dataset	Duration	Glove Types	Ego Views	Exo Views	Both Hands	Scenes
OpenTouch (Song et al., 2025)	5	1	✓ 1	✗ 0	✗	14
TouchAnything (Zhou et al., 2026)	29	1	✓ 1	✗ 0	✓	5
EgoTactile (Zeng et al., 2026)	6	1	✓ 1	✗ 0	✗	1
HumanTouch (Lyu et al., 2026)	100	1	✓ 1	✗ 0	✓	1
ACE-Data-0 (Cao et al., 2026)	16	1	✓ 2	✓ 8	✓	2
EgoFact-Data	100	3	✓ 2	✓ 8	✓	6

Table 1: **Dataset Overview.** EgoFact-Data is a cross-device egocentric-tactile dataset, complementing the limitation of diversity and scale of existing datasets.

3.3 EgoFact-Data Dataset Collection

Although recent egocentric-tactile datasets (Song et al., 2025; Zhou et al., 2026; Zeng et al., 2026; Lyu et al., 2026; Cao et al., 2026) have advanced research on vision-to-tactile estimation, the scale and diversity of available data remain limited. As summarized in Tab. 1, existing datasets are limited to single tactile sensing configuration and lack cross-device data collection. They also face limited recording scale, missing exocentric observations, incomplete handnesses, and insufficient scene and task diversity. To this end, we collect EgoFact-Data, a cross-device egocentric-tactile dataset comprising 100 hours of recordings across 800 clips, collected using three types of tactile gloves in six scenes and covering 200 manipulation tasks. Each recording session lasts approximately 3-15 minutes, during which participants follow predefined task scripts to conduct grasping, picking up, placing, pressing, twisting, tool use, and multi-step complex manipulations. Our data acquisition system uses an ACE-Ego-Head-V02 Lite to capture stereo egocentric video and eight GoPro cameras to record exocentric videos. Full-hand tactile measurements of both hands are collected using three different tactile gloves, including JQ-Glove, Wuji-Glove, and Matrix-Glove. Timestamps are aligned across all modalities at the start of each recording session to establish a common temporal reference. Further details on our data collection and processing are provided in Appendix.

4 Experiments

4.1 Dataset and Evaluation Metrics

We compare our method against baselines on five datasets: OpenTouch (Song et al., 2025), EgoTouch (Zhou et al., 2026), EgoTactile (Zeng et al., 2026), HumanTouch (Lyu et al., 2026), and our EgoFact-Data. We follow the evaluation protocol specified by each dataset, using identical train-test split for comparisons. For ablation study, we conduct all experiments on EgoTouch (Zhou et al., 2026) to maintain a controlled setting and acceptable time cost. Following the practices of baselines (Song et al., 2025; Zhou et al., 2026; Zeng et al., 2026), we evaluate performance using four metrics: frame-level contact-state accuracy (Accuracy), contact region IoU (C-IoU), tactile volume IoU (V-IoU), and mean absolute error (MAE). To ensure consistent evaluation across datasets, all metrics are computed on our unified tactile representation. Accuracy and IoU are reported in percentage, and the reported MAE are denormalized to the real force scale in newtons (N) according to dataset statistics or tactile glove intrinsics. Please refer to Appendix for detailed metric definitions.

4.2 Comparison

Baselines. We compare our method with OpenTouch (Song et al., 2025), TouchAnything (Zhou et al., 2026), and EgoTactile (Zeng et al., 2026) baselines. OpenTouch formulates vision-to-tactile estimation as cross-modal retrieval in a shared embedding space (Radford et al., 2021). TouchAnything takes egocentric and wrist-view videos with additional hand poses to estimate tactile signals. While EgoTactile employs a video diffusion model (Blattmann et al., 2023) to generate tactile heatmaps in pixel space conditioning on egocentric images and segmentation hints. We preserve the native inputs and inference pipelines of all baselines and map their heterogeneous outputs to our unified tactile representation for consistent evaluation.

Results. Our method achieves better overall performance across multiple evaluation tracks. As reported in Tab. 2, on the video-to-tactile track of OpenTouch (Song et al., 2025), our method improves Accuracy from 55.61% to 82.40% and V-IoU from 28.33% to 48.09%, while reducing MAE by 27.7%. On EgoTouch (Zhou et al., 2026) and EgoTactile (Zeng et al., 2026), our method also sub-

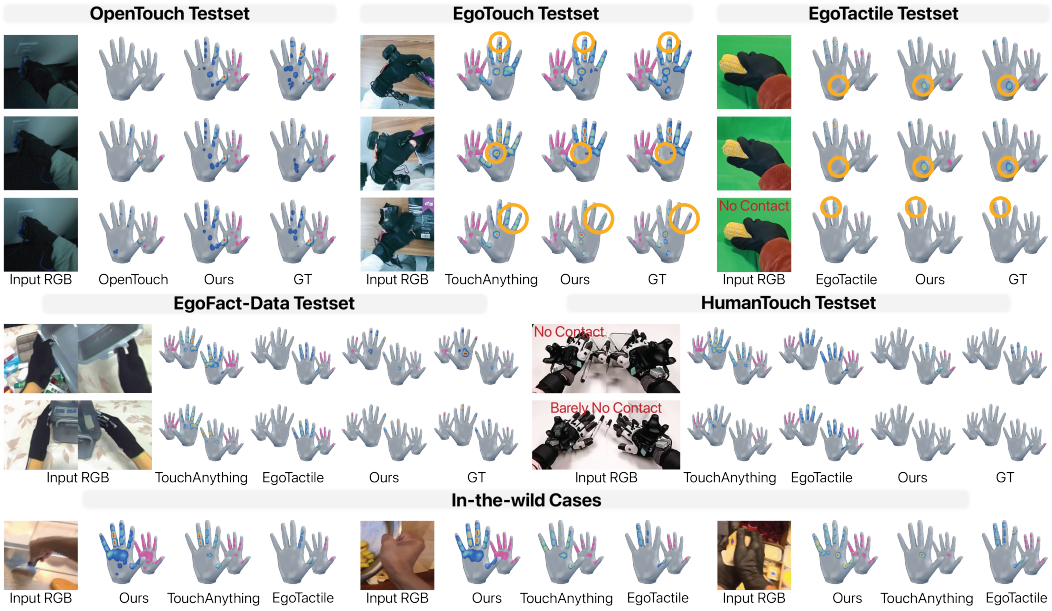

Figure 4: **Qualitative Comparisons** on five benchmarks and in-the-wild cases.

Method	Trained on	Accuracy ↑	C-IoU ↑	V-IoU ↑	MAE ↓
OpenTouch video-to-tactile track					
OpenTouch (Song et al., 2025)	OpenTouch	55.61	33.60	28.33	0.1995
Ours	OpenTouch	**82.40**	**53.15**	**48.09**	**0.1442**
EgoTouch seen object track					
TouchAnything (Zhou et al., 2026)	EgoTouch	87.53	52.52	50.26	**1.4663**
Ours	EgoTouch	**87.61**	**54.85**	**52.56**	1.6186
EgoTouch unseen object track					
TouchAnything (Zhou et al., 2026)	EgoTouch	**84.44**	40.32	39.00	2.5121
Ours	EgoTouch	83.56	**45.73**	**44.01**	**2.3736**
EgoTactile object track					
EgoTactile (Zeng et al., 2026)	EgoTactile	54.20	16.11	23.39	0.4000
Ours	EgoTactile	**79.80**	**49.57**	**45.32**	**0.2273**
EgoTactile subject track					
EgoTactile (Zeng et al., 2026)	EgoTactile	59.97	10.18	26.77	0.3800
Ours	EgoTactile	**67.69**	**40.58**	**38.42**	**0.2229**

Table 2: **In-domain Quantitative Comparisons** on OpenTouch, EgoTouch, and EgoTactile.

		EgoFact-Data scene-held-out		EgoFact-Data task-held-out		HumanTouch
Method	Trained on	Accuracy ↑	MAE ↓	Accuracy ↑	MAE ↓	MAE ↓
OpenTouch (Song et al., 2025)	OpenTouch	72.43	0.0155	47.99	0.0138	0.2251
TouchAnything (Zhou et al., 2026)	EgoTouch	72.33	0.0348	48.29	0.0324	0.3058
EgoTactile (Zeng et al., 2026)	EgoTactile	36.98	0.0153	50.80	0.0228	0.2965
Ours	EgoTouch	72.78	0.0300	49.00	0.0317	0.3544
Ours	EgoFact-Data / HumanTouch	**77.47**	**0.0039**	**53.94**	**0.0053**	**0.1281**

Table 3: **Zero-shot Transfer Comparisons** on EgoFact-Data and HumanTouch.

stantially outperforms the baselines in generalization across unseen subjects and interacting objects. Tab. 3 presents zero-shot performance on the test sets of HumanTouch (Lyu et al., 2026) and our EgoFact-Data, alongside the results of our method trained on each target domain. The results show that our method improves contact-state accuracy by at least 3.14% and reduces MAE by at least 40% relative to the second best results. Under the same zero-shot transfer setting, both trained on EgoTouch (Zhou et al., 2026), our method outperforms TouchAnything (Zhou et al., 2026) on all the metrics of both evaluation tracks of EgoFact-Data. These results support the generalization of our method to unseen scenes and manipulation tasks. Figure 4 further illustrates qualitative comparisons on both in-domain and in-the-wild examples. Across gloved hands, bare hands, and interactions with substantial ambiguity due to occlusions, our method estimates full-hand tactile fields more aligned with tactile sensing and better visual plausibility than baselines.

4.3 Ablation Study

In this section, we first validate the correctness and effectiveness of our proposed representation. We then investigate three key design choices in our method: *(i.)* the visual backbone and feature fusion strategy for tactile feature learning; *(ii.)* the benefits of the force-guided learning objective over the uniformly weighted loss; and *(iii.)* the stability gains from avoiding reliance on hand pose priors supplied by an upstream hand tracker.

Tactile Representations. Tab. 4 compares our unified tactile representation with the direct mapping schemes adopted in prior work (Song et al., 2025; Jeon et al., 2026; Zhang et al., 2026) in terms of 3D spatial smoothness and peak preservation on three distinct datasets (Song et al., 2025; Zhou et al., 2026; Zeng et al., 2026). "Spatial Lap." quantifies local spatial variation in the tactile field over the 3D hand surface, with lower values indicating smoother fields, whereas "Peak Err." measures the peak-amplitude distortion introduced when converting the original tactile signals into the representation. Our representation achieves greater spatial smoothness and lower peak errors than direct mapping. These results show that, beyond unifying heterogeneous tactile data and modeling spatial signal decay over the 3D hand surface, our representation better preserves the intrinsic characteristics of original signals, limiting the loss of salient information.

Repr.	Dataset	Spatial Lap. ↓	Peak Err. ↓
Direct	OpenTouch	37.79	5.22
Ours	OpenTouch	**4.22**	**4.57**
Direct	EgoTouch	36.19	7.13
Ours	EgoTouch	**7.86**	**1.80**
Direct	EgoTactile	2.93	1.15
Ours	EgoTactile	**2.34**	**0.00**

Table 4: **Ablation of Tactile Representations.**

	V-IoU ↑		MAE ↓	
Configuration	Seen	Unseen	Seen	Unseen
HaMeR, final layer	49.59	40.87	1.7409	2.6143
HaMeR, multi-layer	<u>52.01</u>	41.43	**1.5707**	<u>2.4147</u>
DINOv3, final layer	50.02	41.46	1.6988	2.5188
+ HaMeR pose	50.10	41.42	1.6334	2.4539
+ WiLoR pose	49.11	42.51	1.6686	2.4237
Uniform loss	51.35	<u>43.48</u>	1.6656	2.4250
Ours	**52.56**	**44.01**	<u>1.6186</u>	**2.3736**

Table 5: **Ablation of Model Designs on EgoTouch Dataset.**

Visual Backbone and Multi-Level Feature Fusion. To investigate the best practice for tactile feature learning, we compare different visual backbones and ablate our multi-level feature fusion strategy. As shown in Tab. 5, the general-purpose DINOv3 (Siméoni et al., 2025), achieves better tactile estimation performance than expert encoder for hand pose recovery (Pavlakos et al., 2024). Our multi-level feature fusion yields further improvements with both backbones, approximately increasing V-IoU by 3% and reducing MAE by 0.1N. These results suggest that effective tactile feature learning benefits from integrating information across multiple feature levels, rather than relying solely on features of hand kinematics or the final-layer output of the visual backbone. Together, these findings support our choice of a general-purpose visual backbone with multi-level feature fusion.

Hand Pose Priors. Prior works (Zhou et al., 2026; Jeon et al., 2026) incorporate hand-pose priors to guide tactile estimation. To assess the necessity and advantage of such priors, we build variant model architectures that take hand-pose estimates (Pavlakos et al., 2024; Potamias et al., 2025) as additional inputs. As shown in Table 5, incorporating either estimator degrades tactile estimation performance relative to our tracker-free model, particularly in V-IoU. These results indicate that explicit hand-pose priors are unnecessary and can even impair tactile estimation accuracy, supporting our hand-tracker-free architecture.

Force-guided Learning Objective. Finally, we compare our force-guided learning objective, defined in Eq. 5, with the uniformly weighted regression loss. As shown in Table 5, our objective achieves higher V-IoU and lower MAE than uniform weighting. These improvements suggest that force-guided weighting helps the model learn more effectively from tactile data with sensor noises.

4.4 Applications

Benefiting from our vision-to-tactile model and tactile representation defined on the 3D hand surface, we apply the estimated tactile signals to downstream tasks. Here, we demonstrate two representative applications: *(i.)* reconstructing 3D hand-object interactions (HOI) from RGB images; and *(ii.)* incorporating tactile signals as reward functions for imitation learning of manipulation skills in simulation. More details, results, and quantitative evaluations are provided in Appendix.

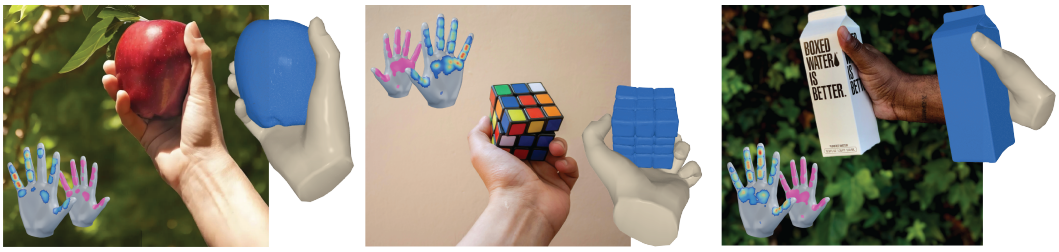

Figure 5: **Application-1.** With dense tactile cues inferred by our method, we can refine the rough 3D hand-object interaction reconstructions to detailed and affordance-correct results.

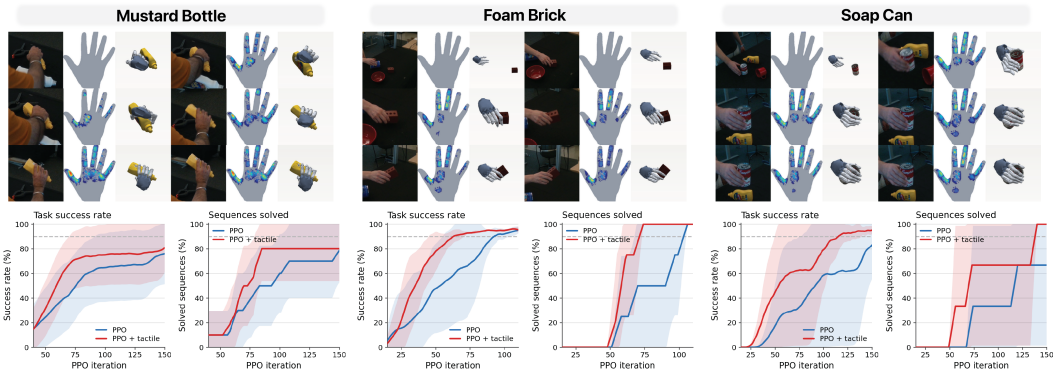

Figure 6: **Application-2.** Additional tactile rewards help to accelerate the imitation process and improve success rate on multiple tasks of DexYCB dataset (Chao et al., 2021b) in simulation.

Tactile-Guided Hand-Object Interaction Reconstruction. Given RGB images, we first obtain a coarse 3D HOI reconstruction using a state-of-the-art method (Aytekin et al., 2026). We then use our estimated contact map and dense tactile field to determine contact regions intensity. A subsequent post-optimization is applied to refine the coarse reconstruction, aligning the contact regions weighted by their estimated tactile intensities. As shown in Figure 5, incorporating such tactile cues produces more accurate 3D HOI reconstructions with finer contact details, reducing penetrations and correcting inconsistent affordances.

Tactile-Aided Dexterous Skill Imitation. Following REGRIND (Feng et al., 2026), we retarget DexYCB (Chao et al., 2021a) demonstrations to Wuji-Hand and learn to imitate them through proximal policy optimization (PPO). We pool the estimated tactile fields into five finger-level reference intensities synchronized with the pose trajectory to form the motion-tracking objective. Fig. 6 illustrates the learning dynamics. At early iteration 60, tactile guidance increases mean completion rate by 20.5%, 14.1%, and 11.8% for three objects, respectively. Moreover, averaged over the final 20 iterations, incorporating tactile guidance improves completion rate from 93.6% to 95.3% for mustard, 85.7% to 86.6% for foam, and 96.6% to 96.8% for can.

5 Conclusion

We presented EgoFact, a unified framework for estimating dense full-hand tactile fields of both hands from egocentric videos, using a proposed unified tactile representation and a tracker-free model with force-guided learning objective. We also introduced cross-device EgoFact-Data with protocols for evaluating generalization to unseen scenes and manipulation tasks. Qualitative and Quantitative experiments demonstrate consistent improvements, with further benefits for 3D hand-object interaction reconstruction and dexterous skill imitation applications. These findings highlight the value of the vision-to-tactile estimation task for physically grounded egocentric intelligence.

REFERENCES

Ayce Idil Aytekin, Helge Rhodin, Rishabh Dabral, and Christian Theobalt. Follow my hold: Hand-object interaction reconstruction through geometric guidance. In *2026 International Conference on 3D Vision (3DV)*, pp. 795–805. IEEE, 2026. 9

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 6

Samarth Brahmbhatt, Cusuh Ham, Charles C Kemp, and James Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8701–8711. IEEE, 2019. 3

Yukang Cao, Haozhe Xie, Beichen Wen, Runmao Yao, Yinghao Liu, Yue Huang, Zhichao Liao, Yunxiang Wang, Haiheng Liu, Xingshun Tian, et al. Ace-data-0: Human-centric ambient capture as embodied data engine. *arXiv preprint arXiv:2607.28625*, 2026. 2, 6

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026. 4

Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. DexYCB: A benchmark for capturing hand grasping of objects. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021a. 9

Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9040–9049. IEEE, 2021b. 3, 9

Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12943–12954. IEEE, 2023. 3

Yunhai Feng, Natalie Leung, Jiaxuan Wang, Lujie Yang, Haozhi Qi, and Preston Culbertson. A minimalist retargeting-guided reinforcement learning recipe for dexterous manipulation. *arXiv preprint arXiv:2607.11874*, 2026. 9

Patrick Grady, Chengcheng Tang, Samarth Brahmbhatt, Christopher D Twigg, Chengde Wan, James Hays, and Charles C Kemp. Pressurevision: estimating hand pressure from a single rgb image. In *European Conference on Computer Vision*, pp. 328–345. Springer, 2022. 3

Patrick Grady, Jeremy A Collins, Chengcheng Tang, Christopher D Twigg, Kunal Aneja, James Hays, and Charles C Kemp. Pressurevision++: Estimating fingertip pressure from diverse rgb images. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 8683–8693. IEEE, 2024. 3

Andela Ilic, Rachel Schuchert, Yijing Jiang, and Christian Holz. Egophi: Estimating 3d hand-object contact and force from egocentric vision. In *European Conference on Computer Vision*, pp. 410–428. Springer, 2026. 3

Subin Jeon, Byungjun Kim, and Hanbyul Joo. Hope: Hand-object pressure estimation from monocular videos. *arXiv preprint arXiv:2608.06192*, 2026. 2, 3, 4, 5, 8

Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20981–20990. IEEE, 2022. 3

Chuqiao Lyu, Wenbo Ding, Tianxing Chen, and Qi Xiong. Humantouch: A multimodal system for scalable human-hand tactile acquisition, 2026. SparkLab@Xspark AI. 2, 3, 6, 7

Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9826–9836. IEEE, 2024. 5, 8

Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12242–12254. IEEE, 2025. 4, 5, 8

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021. 6

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12159–12168. IEEE, 2021. 5

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016. 4

Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022. 2, 3

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 5, 8

Yuxin Ray Song, Jinzhou Li, Rao Fu, Devin Murphy, Kaichen Zhou, Rishi Shiv, Yaqi Li, Haoyu Xiong, Crystal Elaine Owens, Yilun Du, et al. Opentouch: Bringing full-hand touch to real-world interaction. *arXiv preprint arXiv:2512.16842*, 2025. 2, 3, 4, 6, 7, 8

Markku Suomalainen, Yiannis Karayiannidis, and Ville Kyrki. A survey of robot manipulation in contact. *Robotics and Autonomous Systems*, 156:104224, 2022. 1

Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *European conference on computer vision*, pp. 581–600. Springer, 2020. 3

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in neural information processing systems*, 35:38571–38584, 2022. 4

Akihiko Yamaguchi and Christopher G Atkeson. Recent progress in tactile sensing and sensors for robotic manipulation: can we turn tactile sensing into vision? *Advanced Robotics*, 33(14): 661–673, 2019. 1

Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu. Oakink: A large-scale knowledge repository for understanding hand-object interaction. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20921–20930. IEEE, 2022. 3

Yuan Zeng, Yujia Shi, Tiao Tan, Xingting Li, Yaqi Qin, Zongqing Lu, Wenming Yang, Jing-Hao Xue, and Qingmin Liao. Egotactile: Learning grasp pressure for everyday objects from egocentric video. *arXiv preprint arXiv:2606.09243*, 2026. 2, 3, 4, 6, 7, 8

Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. Hawor: World-space hand motion reconstruction from egocentric videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1805–1815. IEEE, 2025. 5

Wenkang Zhang, Chengbo Yuan, Zicheng Zhang, Zhengxue Cheng, and Yang Gao. Egotac: In-the-wild tactile prediction from egocentric vision. *arXiv preprint arXiv:2608.15060*, 2026. 2, 3, 4, 5, 8

Chengfeng Zhao, Juze Zhang, Jiashen Du, Ziwei Shan, Junye Wang, Jingyi Yu, Jingya Wang, and Lan Xu. I'm hoi: Inertia-aware monocular capture of 3d human-object interactions. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 729–741. IEEE, 2024. 3

Yiming Zhao, Taein Kwon, Paul Streli, Marc Pollefeys, and Christian Holz. Egopressure: A dataset for hand pressure and pose estimation in egocentric vision. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27727–27738. IEEE, 2025. 2, 3

Jianyi Zhou, Ziteng Gao, Feiyang Hong, Zirui Liu, Guannan Zhang, Weisheng Dai, Ruichen Zhen, Chuqiao Lyu, Haotian Wu, Yinian Mao, et al. Touchanything: A dataset and framework for bimanual tactile estimation from egocentric video. *arXiv preprint arXiv:2605.13083*, 2026. 2, 3, 5, 6, 7, 8