arXiv:2610.05180v2 [cs.RO] 14 Oct 2026

VICON: Visual-Inertial-Contact based Hand-Object Tracking for Manipulation Datasets

Youbeen Jeon[1], Uiseong Shin[1], Hwanchul La[1], Jaeseong Kang[1], Hyelim Choi[2] and Yongseok Lee[†]

***Abstract*— Learning dexterous manipulation benefits from human demonstration datasets that capture diverse and natural hand-object interactions. In particular, contact points and forces provide supervision on where and how strongly to interact, which cannot be fully captured by motion trajectories alone. However, methods for jointly capturing hand and object motion, contact points, and forces remain limited. Moreover, severe occlusion from hand-object interaction challenges accurate tracking of both hands and objects. To address these limitations, we present a Visual-Inertial-CONtact based hand-object tracking (VICON) framework. It holistically captures both hand and object motion along with contact information during manipulation, even under severe occlusion. First, we adopt a visual-inertial glove [1] and an RGB-D camera for accurate hand tracking, and redesign the glove to incorporate contact sensing. Specifically, force-sensitive resistors (FSRs) are placed on the glove based on human grasp frequency to synchronously record contact states and calibrated normal forces. Second, without requiring pre-existing CAD models, we estimate object poses using RGB-D images and a mesh reconstructed from a monocular video. We propose factor-graph-based object trajectory estimation that fuses object-pose estimates weighted by visibility under hand-object occlusion, FSR measurements, and a hand-motion prior. Across 40 motion-capture sessions with five objects, VICON achieves a 2.5% failed-frame rate compared with 50.9–64.6% for the baselines, with median errors of 3.9 mm and 3.0° under occlusion. Using VICON, we construct a dataset containing synchronized hand-object motion, contact points, and normal forces, and will publicly release an expanded dataset covering 10 object categories at `https://github.com/VICON-dataset/dataset`.**

I. Introduction

Dexterous manipulation is a fundamental capability for humanoid robots operating in human environments, where everyday tasks require coordinated hand motion and continuous physical interaction with objects. Learning such intricate behaviors heavily relies on rich demonstrations. However, collecting demonstrations via teleoperation is challenging; differences in kinematics and workspace between human and robot hands make real-time control unintuitive, severely limiting data collection efficiency and degrading motion quality. In contrast, directly capturing human demonstrations bypasses these difficulties, naturally preserving human dexterity and task variability across diverse objects and environments [2], [3]. Yet, to fully realize this potential, hand-object interactions must be measured reliably. Rich demonstration data requires more than hand-object motion alone, since successful dexterous manipulation fundamentally depends on where contact occurs and how much force is applied. Such information is particularly important for learning contact-rich manipulation policies [4] and emerging force-aware vision-language-action models [5]. Demonstration data should therefore capture contact points and forces alongside hand and object motion. However, many existing datasets provide detailed hand-object geometry and contact annotations without directly measuring the applied forces [6], [7]. Although recent tactile datasets and sensing systems provide richer contact information [8]–[10], simultaneously capturing hand-object poses along with contact locations and forces remains challenging within a compact setup. In particular, achieving accurate ground truth often requires multi-view camera systems, restricting data collection to dedicated environments.

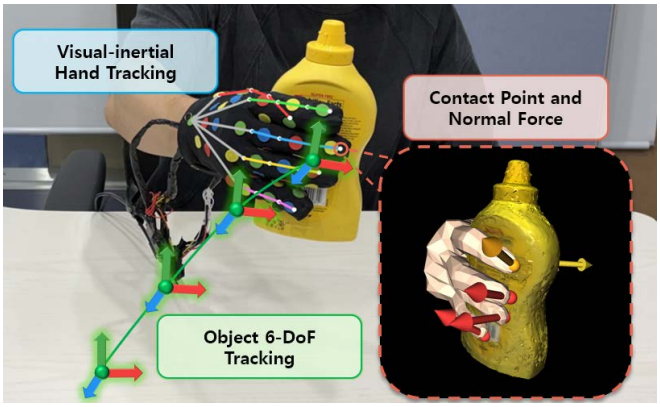


Fig. 1. VICON records hand motion, object pose, contact points, and forces using a visual-inertial glove [1] with integrated FSRs and a single RGB-D camera. The glove tracks the hand skeleton, and our offline refinement maintains robust and accurate object 6-DoF tracking even under severe occlusion. Inset: object-frame contact points, with arrows colored by calibrated normal force magnitude.

With a compact setup, occlusion presents another major challenge for reliable hand-object tracking. Under severe occlusion, even state-of-the-art vision-based pose estimators can become unreliable or fail entirely. Although existing works have attempted to mitigate occlusion in hand pose estimation [11] and hand-object tracking [12]–[14], these approaches still rely heavily on visual cues, which can become unreliable under severe or prolonged occlusion. Recent work has explored using hand motion as an additional constraint for object tracking [15], but maintaining robust tracking when visual evidence is largely unavailable remains challenging.

Wearable hand tracking directly captures hand kinematics, providing reliable motion estimates where vision-only

Youbeen Jeon, Uiseong Shin, Hwanchul La, Jaeseong Kang, and Yongseok Lee are with DGIST, Daegu 42988, Republic of Korea (e-mail: yoobin1liyfcb@gmail.com; tlsdmltjd52@gmail.com; hc.la@dgist.ac.kr; rkd020320@gmail.com; yslee@dgist.ac.kr). Hyelim Choi is with Seoul National University, Republic of Korea (e-mail: helmchoi@snu.ac.kr).

[1]These authors contributed equally to this work as first authors.

[2]Second author.

[†]Corresponding author: Yongseok Lee (e-mail: yslee@dgist.ac.kr).

estimation fails. In particular, visual-inertial skeletal tracking (VIST) [1] combines inertial measurements with visual marker observations to robustly track hand motion under severe occlusions. Because hand and object trajectories are tightly coupled during manipulation, this robust hand motion can serve as a prior to guide object pose estimation when visual object cues become unreliable. However, this kinematic constraint is physically valid only when the hand and object are in direct contact. Contact sensing therefore serves as a crucial gate to determine when the hand-motion prior should be applied. To address this, we integrate force-sensitive resistors (FSRs) directly onto the glove, providing occlusion-invariant contact detection while simultaneously yielding calibrated normal force measurements.

Building on these components, we present **VICON** (Fig. 1), a multimodal tracking framework built on the VIST glove [1] integrated with eight FSRs, together with a single RGB-D camera. An offline factor graph fuses object-pose estimates from FoundationPose [16], weighted by object visibility, with a contact-gated hand-motion prior to refine object trajectories under occlusion. Using the refined trajectories, contact points are located on the object surface, which are paired with calibrated FSR forces. The main contributions of our work are:

1) **Visual-inertial hand tracking system with contact sensing.**
We integrate eight FSRs into the VIST glove [1], positioned according to human grasp contact frequency [17]. Together with a single RGB-D camera, the resulting **VICON-glove** synchronously records hand motion, contact states, and calibrated normal contact forces.
2) **Robust hand-object tracking algorithm under severe occlusion.**
We fuse a contact-gated hand-motion prior with visibility-weighted object pose estimates to refine object trajectories and maintain robust tracking even under severe occlusion.
3) **Scalable contact-rich dataset collection for custom objects.**
We present a pipeline that reconstructs object geometry from monocular video and captures synchronized hand motion, object pose, object-frame contact points, and calibrated normal forces using only a VICON-glove and a single RGB-D camera. We will publicly release an expanded dataset covering 10 object categories.

II. Related Works

A. Vision-Based Hand and Object Pose Estimation

Vision-based methods have been widely studied for recovering hand and object poses during manipulation from RGB or RGB-D observations. For object pose estimation, MegaPose [18] estimates the 6-DoF pose of novel objects using rendered CAD views for coarse pose selection and iterative refinement. FoundationPose [16] supports both estimation and tracking with CAD models or neural object representations learned from reference RGB-D images, although occlusion degrades the visual cues needed for reliable alignment. A second line of work jointly estimates hand and object poses by aligning hand-object models with RGB-D observations using proximity-based contact and occlusion constraints [12]. HOLD [14] reconstructs both from monocular video using neural implicit surfaces and contact constraints without an object template. More recently, ComPose [15] learns from monocular RGB video to weight hand joints and combine hand- and object-derived motion cues for object tracking. Despite these advances, image-derived geometry and motion consistency may not reliably distinguish near-contact state from actual contact at an occluded surface, and none of these methods directly measures contact forces.

TABLE I
Comparison of our dataset with existing interaction datasets

Dataset	Hand pose	Object pose	Contact format	Force/ tactile	Setup
EgoDex [2]	✓	–	–	–	Vision Pro
ContactPose [6]	✓	✓	Object surface map	–	3 RGB-D, thermal, mocap
ARCTIC [7]	✓	✓	Object surface map	–	8+1 RGB, mocap
OpenTouch [9]	✓	–	Hand map	Pressure	Aria glasses, tracking and tactile gloves
TactiDex [10]	✓	✓	Hand map	Pressure	Mocap, tracking and tactile gloves
ForceVLA-Data [5]	–	–	–	TCP wrench (est.)	Robot arm, 2 RGB-D
VICON (ours)	✓	✓	Object-frame 3D points	Normal force (N)	1 RGB-D, 1 VICON-glove (8 FSRs)

✓: provided; –: not provided. Contact format: stored representation of contact (ContactPose: thermal imprint after a static grasp; ARCTIC: derived from geometry). Force/tactile: recorded signal type.

B. Hand-Tracking Gloves and Contact Sensing

Sensorized gloves have been developed to capture hand motion and tactile interactions during manipulation. VIST [1] fuses glove-mounted IMUs with visual markers under hand kinematic constraints, using inertial propagation and visual corrections to maintain hand tracking through occlusion. A markerless approach uses inertial pose priors to infer image keypoints and their uncertainty and combines the learned keypoint likelihoods with inertial observations in a factor graph [19]. FSGlove [20] uses IMUs placed on the glove, whose alignment is calibrated from reference poses and fits a personalized hand shape using predefined finger contact constraints. However, an external dorsal tracker is additionally used to recover the hand's global pose. For contact sensing, STAG [8] records spatial patterns of normal force using 548 piezoresistive sensors distributed over the hand, and learns to identify objects and estimate their weights from these tactile signatures. However, neither kinematic tracking nor tactile sensing is sufficient on its own; kinematic tracking cannot measure applied force, while tactile sensing requires

hand-object kinematics to locate contacts on the object surface. Despite this clear complementarity, prior works rarely integrate both modalities in a single wearable system.

C. Hand-Object Interaction Datasets for Manipulation

Table I compares existing datasets in terms of pose annotations, contact and force signals, and recording setups. EgoDex [2] pairs large-scale egocentric video with hand poses but has no object pose, contact, or force. ContactPose [6] maps thermal imprints after static grasps onto object meshes alongside hand and object poses. ARCTIC [7] derives contact from reconstructed hand-object geometry during bimanual manipulation of articulated objects, requiring a costly multi-camera motion-capture studio.

Tactile datasets pair measured pressure with hand motion. OpenTouch [9] combines full-hand pressure, hand tracking, and egocentric video, but provides no object poses. More recently, TactiDex [10] combines dense tactile gloves with hand-object motion capture, using pressure-gated proximity and tactile-constrained optimization, but requires a motion-capture workspace.

Force-aware robot datasets record external loads for contact-rich policy learning. ForceVLA [5] synchronizes images, proprioception, and estimated six-axis tool-center-point (TCP) wrenches in ForceVLA-Data across five contact-rich tasks and fuses them in a vision-language-action policy. However, these wrenches describe net end-effector loads without resolving individual human hand-object contacts.

In contrast to these systems, VICON enables scalable collection of contact-rich demonstrations. Specifically, it is low-cost and wearable, since the VICON-glove builds on the VIST glove, whose material cost is about $100 and weight is 52–55 g [1], and adds only eight FSRs. Moreover, it requires only one RGB-D camera and no motion-capture studio, so recording is not restricted to a dedicated capture space. In addition, it provides accurate data even under severe occlusion, as the glove tracks the hand robustly and the FSRs measure contact directly (Sec. V-C). Finally, it pairs object-frame contact points with calibrated normal forces, a feature absent from prior datasets in Table I. As the volume of demonstration data becomes increasingly important for learning manipulation, this combination of low acquisition cost and high accuracy is a key advantage over costly studio-based systems such as ARCTIC [7].

III. Hardware System

The hardware consists of the VICON-glove (Fig. 2) and one Intel RealSense D435i camera. The glove is built on VIST [1], which incorporates IMUs and visual markers, and is augmented with eight FSRs. The camera is positioned beside the subject at roughly eye level, providing a side view of the workspace; depth is computed from the infrared pair with Fast-FoundationStereo [21] at about 30 Hz.

A. Hand-Tracking Glove

The glove consists of two layers, an inner glove carrying the IMUs and an outer glove carrying the markers. Fourteen

Fig. 2. Multi-layer VICON-glove. (a) Placement of the 14 IMUs in the inner glove. (b) Outer glove with the visual markers. (c) Hand contact-frequency map [17], averaged over the mechanic and housemaid tasks (values in %), used to choose the FSR sites. (d) Palmar view of the eight FSRs.

IMUs are mounted on the inner glove: one on the dorsum of the hand and thirteen along the fingers, with no IMUs on the distal phalanges of the ring and little fingers (Fig. 2(a)). The outer glove holds 80 markers in four colors (Fig. 2(b)), and visual observations of these markers are used to correct drift in the inertial estimates.

B. FSR Placement

We determine the FSR placement based on a contact-frequency analysis of human grasps(Fig. 2(c)) [17]. Their study classifies grasps into power, precision, and intermediate grasps and analyzes contact frequencies during mechanic and housemaid tasks. The results show that the fingertips exhibit the highest contact frequencies, while frequent contact is also observed at the thenar eminence and distal palm. Based on these observations, we place a sensor on each of the five fingertips, one on the thenar eminence, and two on the distal palm over the second- and fifth-metacarpal regions (Fig. 2(d)). For the ulnar-side distal-palm sensor, the analysis reports contact frequency as a single region spanning the third to fifth metacarpals, without distinguishing the areas beneath the ring and little fingers [17]. We therefore place the FSR near the little-finger metacarpophalangeal(MCP) joint on the palmar side to provide broader spatial coverage within this region.

IV. Visibility- and Contact-Guided Object Pose Refinement

A. Notation and Overview

Throughout, $\mathbf{T}_{AB} \in SE(3)$ denotes the pose of frame B expressed in frame A. We use the stationary camera frame C, the hand frame H (the palm frame), and the object frame O attached to the mesh. We estimate the object poses $\mathbf{X}_k = \mathbf{T}_{CO,k} = (\mathbf{R}_k, \mathbf{p}_k)$ over the frames k of a session, while the hand poses $\mathbf{H}_k = \mathbf{T}_{CH,k}$ and the centers $\mathbf{s}^C_{i,k}$ of the eight FSRs

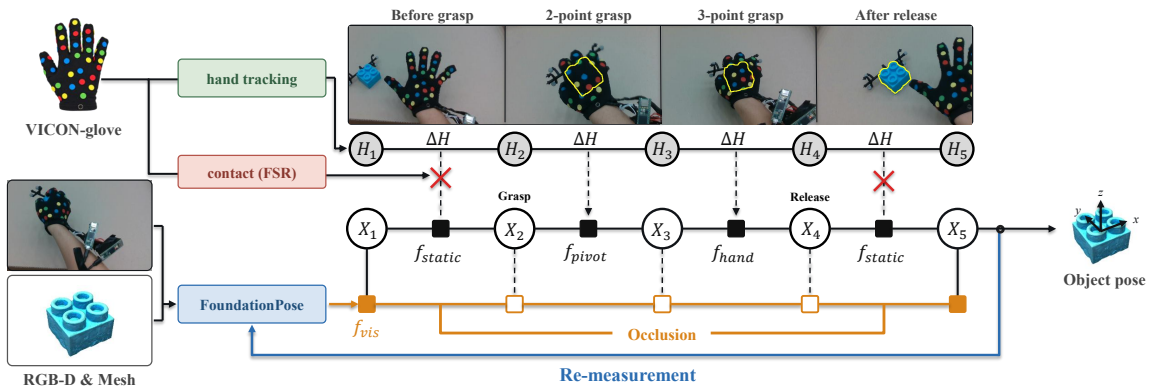


Fig. 3. VICON on a recorded session. The VICON-glove provides hand tracking (poses $\mathbf{H}_1$–$\mathbf{H}_5$, increments ΔH) and contact (FSR); the RGB-D camera feeds FoundationPose, whose candidates enter as f_{vis} (filled: accepted; hollow: rejected under occlusion). Between object poses X_k, contact selects f_{stat} (visible, after release; red crosses: rejected hand increments), f_{pivot} (2-point grasp), or f_{hand} (3-point grasp). Re-measurement feeds the refined poses back to FoundationPose.

($i = 1, \ldots, 8$), where a superscript denotes the frame in which a point is expressed, are provided by the glove and treated as fixed inputs during optimization.

Offline refinement takes as input the initial FoundationPose trajectory $\{\widetilde{\mathbf{X}}_k\}$, the recorded RGB-D frames, the hand trajectory, and the FSR streams (Fig. 3). FoundationPose is initialized from a *seed pose* ($\widetilde{\mathbf{X}}_k$ at first, the refined trajectory later) and returns a *candidate* $\mathbf{Z}_k$. Candidates are evaluated by visibility (Sec. IV-B). Contact extraction then alternates with refinement (Sec. IV-C): each round begins with a *re-measurement*, re-running FoundationPose bi-directionally from the seed poses (pose-refinement network only), and ends in a factor graph fusing the visibility-weighted candidates with f_{hand}, f_{pivot}, and f_{stat} (Sec. IV-D). Contact points and forces follow from the refined poses (Sec. IV-E).

B. Visual Evidence and Visibility Weighting

Since FoundationPose keeps outputting poses under occlusion even when little of the object is visible, we score each candidate pose $\mathbf{Z}_k$ by comparing the rendered mesh depth D_O with the observed depth D. Let Ω denote the set of pixels with $D_O(u), D(u) > 0.001$, and let $\delta(u) = D(u) - D_O(u)$, with depths in meters. We define

$$v_d(\mathbf{Z}_k) = \frac{1}{|\Omega|} \sum_{u \in \Omega} \mathbf{1}[|\delta(u)| < 0.010], \tag{1}$$

$$v_{\text{far}}(\mathbf{Z}_k) = \frac{1}{|\Omega|} \sum_{u \in \Omega} \mathbf{1}[\delta(u) > 0.020], \tag{2}$$

where $\mathbf{1}[\cdot]$ is the indicator function. The visibility v_d falls under either occlusion or pose error, reducing the measurement weight, while v_{far} flags candidates floating in front of the measured surface.

Before re-measurement, the tracked hand skeleton and forearm are projected as capsules to form a hand mask; depths inside it more than 20 mm in front of the seed rendering are removed, with a dilated margin, from the depth image passed to FoundationPose; the returned candidates are scored against the original depth.

TABLE II
NOMINAL VISUAL-MEASUREMENT UNCERTAINTY AS A FUNCTION OF VISIBILITY

v_d	0.10	0.15	0.30	0.50	0.70	0.90	1.0
σ_p (mm)	6	5	3	3	2	1.2	0.8
σ_r (deg)	180	180	8	3	2	1	0.8

Nominal standard deviations of the visual measurements before the per-candidate adjustments of Sec. IV-B; 180° denotes a weak but finite rotational weight.

A candidate is accepted only if $|\Omega| \geq 200$, $v_d \geq 0.10$, it does not penetrate the table by more than 25 mm, and v_{far} is below a threshold that defaults to 0.15 and is adapted per session from high-visibility candidates. A candidate must then persist over consecutive frames without jumping from the previous accepted pose, agree with the anchors of its contact-free interval, and, during grasp, stay near the propagated pose $\mathbf{B}_k$.

For accepted measurements, the deviations σ_p and σ_r (Table II) are interpolated between visibility knots with a cubic smoothstep and adjusted per candidate: σ_p grows with displacement from the seed pose, penetration of the table, and, during a grasp, the distance the hand travels in the tracker's 100 ms lag, while heavy hand coverage raises the floor of both deviations. Below a visibility threshold (0.50 by default, lower for temporally stable candidates), only the depth-supported tilt is retained with its own deviation and the twist about that axis keeps a weak but finite weight. These gates, masks, and covariances define the *visibility weighting* in f_{vis}.

C. Tactile Contact Timing and Hand-Motion Prior

FSR activation and FSR–object proximity determine whether the *hand-motion prior* is applied at each frame. Each FSR signal is sampled at 200 Hz, low-pass filtered at 20 Hz, and converted to a per-frame binary activation $a_{i,k} \in \{0, 1\}$ using a Schmitt trigger with a short hold time. Let $d_{i,k}$ be the unsigned distance from the center of FSR i to the mesh at a reference pose, which is kept fixed while the contact states are extracted. FSR i is *recognized* at frame k when

$a_{i,k} = 1$ and $d_{i,k} \leq 0.030$; writing $\rho_{i,k} = a_{i,k}\mathbf{1}[d_{i,k} \leq 0.030]$, we declare a grasp when at least three FSRs are recognized,

$$c_k = \mathbf{1}\left[\sum_{i=1}^{8} \rho_{i,k} \geq 3\right]. \tag{3}$$

Because the reference poses carry their own error, contact states are extracted twice. The first pass registers a contact when an FSR activates within 100 mm of the initial poses and keeps it until it deactivates, so that intervals survive frames with inaccurate initial poses; a first refinement then yields a trajectory accurate enough to serve as reference. In the second pass, contact states are re-extracted from that trajectory using (3), and a second refinement yields all results below. We call the criterion (3) *tactile contact timing* (contact timing for short); it only gates when the prior applies and does not model force magnitude or force closure. Frames with $c_k = 1$ define grasp intervals, in which the object moves rigidly with the hand (f_{hand}). Adjacent runs of at least five frames with exactly two recognized fingertip contacts (2-point grasp), which occur during grasp formation, release, or weakened contact, are also propagated with the hand, with the prior relaxed about the line through the two contacts (f_{pivot}).

Alongside this contact gate, we maintain a propagated pose $\mathbf{B}_k$. We call accepted measurements with $v_d \geq 0.7$ *anchors*; $\mathbf{B}_k$ is initialized from the first anchor and, while the object is not grasped, updated at every anchor. During a grasp with valid hand tracking (a *usable hand increment*), it is propagated with the hand, $\mathbf{B}_k = \mathbf{H}_k\mathbf{H}_{k-1}^{-1}\mathbf{B}_{k-1}$, and remains fixed when neither a visual update nor a usable hand increment is available. Across *visual gaps*, runs of frames without accepted measurements, the propagated pose is reconciled with the anchors on either side: agreeing anchors indicate a stationary object, while a mismatch is distributed over the hand-propagated frames. A grasp interval without any anchor is assigned synthetic poses $\bar{\mathbf{Z}}_k$ from the propagated trajectory, with large nominal deviations of 80 mm and 45°.

D. Object-Pose Factor Graph

These components are combined in a single factor graph over $\{\mathbf{X}_k\}$, while the hand poses, contact states, and object geometry are kept fixed. Let $\mathbf{M}_k = \mathbf{B}_{k-1}^{-1}\mathbf{B}_k$ denote the increment of the propagated pose from frame $k-1$ to k, expressed in the object frame. It equals the hand motion, plus the anchor correction where it spans a visual gap. The residuals of the visual, synthetic, hand-motion, stationary, and grasp-link factors are defined as

$$r_k^{\text{vis}} = \text{Log}(\mathbf{Z}_k^{-1}\mathbf{X}_k), \tag{4}$$
$$r_k^{\text{syn}} = \text{Log}(\bar{\mathbf{Z}}_k^{-1}\mathbf{X}_k), \tag{5}$$
$$r_k^{\text{hand}} = \text{Log}(\mathbf{M}_k^{-1}\mathbf{X}_{k-1}^{-1}\mathbf{X}_k), \tag{6}$$
$$r_k^{\text{stat}} = \text{Log}(\mathbf{X}_{k-1}^{-1}\mathbf{X}_k), \tag{7}$$
$$r_{kl}^{\text{grasp}} = \text{Log}(\mathbf{X}_k^{-1}\mathbf{H}_k\mathbf{H}_l^{-1}\mathbf{X}_l), \tag{8}$$

where $\text{Log} : SE(3) \to \mathbb{R}^6$ uses rotation-first tangent coordinates, and the grasp-link residual enforces a constant hand-to-object transform $\mathbf{H}_k^{-1}\mathbf{X}_k$ across linked frames. With $\|r\|_A = (r^T A r)^{1/2}$ and the Huber loss ϕ (threshold 1.345), we minimize

$$\begin{aligned}
&\sum_{k\in\mathcal{V}} \phi\left(\|r_k^{\text{vis}}\|_{\Sigma_k^{-1}}\right) + \sum_{k\in\bar{\mathcal{V}}} \phi\left(\|r_k^{\text{syn}}\|_{\Lambda_k^{-1}}\right) \\
&+ \tfrac{1}{2}\sum_{k\in\mathcal{H}} \|r_k^{\text{hand}}\|^2_{Q_k^{-1}} + \tfrac{1}{2}\sum_{k\in\mathcal{W}} \|r_k^{\text{stat}}\|^2_{W_k^{-1}} \\
&+ \tfrac{1}{2}\sum_{(k,l)\in\mathcal{A}} \|r_{kl}^{\text{grasp}}\|^2_{P_{kl}^{-1}} \\
&+ \sum_{(k,l)\in\mathcal{G}} \phi\left(\|r_{kl}^{\text{grasp}}\|_{U_{kl}^{-1}}\right)
\end{aligned} \tag{9}$$

The sets $\mathcal{V}$, $\bar{\mathcal{V}}$, $\mathcal{H}$, $\mathcal{W}$, $\mathcal{A}$, and $\mathcal{G}$ index accepted measurements, synthetic poses, usable hand increments, stationary pairs $(k-1,k)$, adjacent grasp pairs without a usable hand increment, and the long-range grasp links described below.

Except for the pivot relaxation below, all 6×6 covariances are block-diagonal in rotation and translation. The visual covariance Σ_k uses the adjusted per-candidate deviations of Sec. IV-B, and the synthetic-pose covariance Λ_k is isotropic within each block. The base hand-motion covariance Q_k^0, stationary covariance W_k, and adjacent grasp-link covariance P_{kl} are diagonal, with translation/rotation deviations of 0.55 mm/0.2°, 0.01 mm/0.03°, and 1.5 mm/1.5° per nominal frame interval Δt_{nom}, rescaled by $\Delta t_k/\Delta t_{\text{nom}}$, visual disagreement, occlusion, and hand-tracking quality. For grasps of at least ten frames, long-range links connect each grasp frame to the anchor frames on either side, with an 8° rotational deviation in U_{kl} and translation left effectively unconstrained. For each pair in a 2-point grasp, let $\mathbf{s}_{j,k-1}^O = \mathbf{B}_{k-1}^{-1}\mathbf{s}_{j,k-1}^C$, $j \in \{a,b\}$, denote the two recognized FSR centers in the object frame, $\mathbf{n}_{k-1}$ the unit vector along the line connecting them, and μ_{k-1} its midpoint. The rigid hand-motion prior is relaxed about this line through a rank-one covariance update,

$$Q_k^{\text{piv}} = Q_k^0 + (\sigma_{\text{free}}\alpha_k)^2\,\mathbf{u}_k\mathbf{u}_k^T, \tag{10}$$

with $\mathbf{u}_k = [\mathbf{n}_{k-1}; \mu_{k-1}\times\mathbf{n}_{k-1}]$ the screw axis of rotation about the contact line, expressed in the object frame as in (6), $\alpha_k = \max(\Delta t_k/\Delta t_{\text{nom}}, 0.2)$, and $\sigma_{\text{free}} = 15°$ per nominal frame interval. It applies only to pairs with a valid axis and $\Delta t_k \leq 0.25$ s, lowering the penalty on rotation about the contact line and its coupled translation. In (9), $Q_k = Q_k^{\text{piv}}$ for these pairs (f_{pivot}) and $Q_k = Q_k^0$ otherwise (f_{hand}).

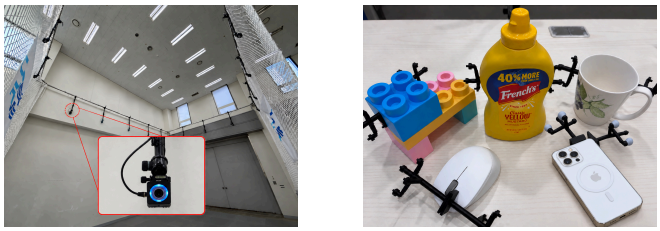
The graph is initialized with the propagated poses and is solved in two stages using Levenberg–Marquardt in GTSAM. The first stage uses a synthetic-pose rotational deviation of 5°, and the second stage uses the nominal value of 45°.

E. Contact and Force Annotation

The 12-bit ADC readings from the FSRs are first converted into a dimensionless response and then mapped to normal force by a per-FSR calibration,

$$x = \frac{\text{ADC}}{\max(4095 - \text{ADC}, 1)}, \qquad F_i = f_i(x). \tag{11}$$

Because the datasheet specifies ±6% part-to-part repeatability [22], each FSR was calibrated separately against a digital scale over slow press–release cycles up to 61 N. Each



(a) (b)

Fig. 4. Experimental setup. (a) Motion-capture room; the inset shows one of the motion-capture cameras. (b) The five objects: block, mouse, mug, mustard bottle, and phone.

f_i is an affine fit $a_i x + b_i$ or a power law $a_i x^{n_i}$, whichever gives the higher R^2, joined linearly to the origin below the lowest calibrated point. The fits yield $R^2 = 0.83$–0.99 and mean absolute errors of 0.74–2.87 N overall and 0.68–2.50 N below 20 N, where most recorded forces lie (Sec. V-E). No FSR saturated within the calibrated range.

For contact annotation, each recognized FSR is associated with the nearest point on the object mesh $\mathcal{M}_O$ at the refined pose:

$$\mathbf{q}^O_{i,k} \in \arg\min_{\mathbf{q}\in\mathcal{M}_O} \left\|\mathbf{X}_k\mathbf{q} - \mathbf{s}^C_{i,k}\right\|_2, \qquad \mathbf{q}^C_{i,k} = \mathbf{X}_k\mathbf{q}^O_{i,k}. \tag{12}$$

Each record stores $\mathbf{q}^O_{i,k}$ and $\mathbf{q}^C_{i,k}$, the FSR–surface distance signed negative inside the mesh, and the synchronized force $F_{i,k} = f_i(x_{i,k})$.

V. Experiments and Results

A. Recording Setup and Object Models

Four subjects, each wearing a single glove, recorded 71 sessions with five rigid objects: a block, a mouse, a mug, a mustard bottle, and a phone (Fig. 4(b)), and each session consisted of approach, grasp, slow transport, and placement. Forty sessions (i.e., eight per object) were captured in a motion-capture room for ground truth (Fig. 4(a)); the other 31 were recorded without motion capture. The object meshes were obtained from monocular video via COLMAP [23], 3DGS [24], and SuGaR [25], with a ChArUco board setting the metric scale and unobserved surfaces cleaned away; all evaluated object pose trackers use the same meshes. Validating on YCB objects [26], the reconstructions deviate from the scans by only about 1 mm on average and yield closely matching FoundationPose trajectories, so we use the reconstructions throughout.

B. Ground Truth and Metrics

To validate the accuracy of our framework, we evaluate object-pose accuracy, using the motion-capture results as our ground truth (GT). The motion-capture and camera coordinate frames were aligned and the offsets of the motion-capture markers with respect to the mesh were calibrated through static scenes.

All trackers produce predictions on the same 47,222 frames in which at least four motion-capture markers are visible with a fit residual within 2 mm. To stratify them by occlusion we evaluate (1) at the GT pose, using the full mesh and the unmasked depth, and call the result v_d^{GT}: a frame is occluded

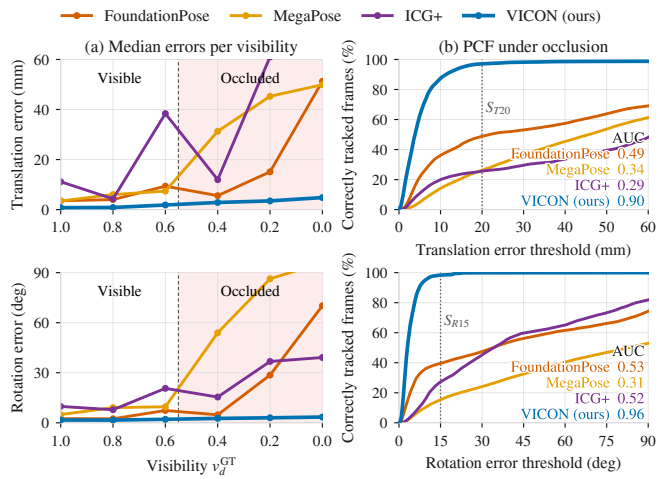


Fig. 5. Object-pose errors over the 40 motion-capture sessions (top: translation; bottom: symmetry-aware rotation). (a) Median error of each tracker binned by the visibility v_d^{GT}; shading marks the occluded range. (b) PCF on the 24,285 occluded frames as a function of the error threshold; dotted lines mark the 20 mm and 15° thresholds used for S_{T20} and S_{R15}, and the numbers give the AUC. Medians are used in (a) because diverging baseline errors would dominate a mean.

when $v_d^{\mathrm{GT}} < 0.55$ and visible otherwise. The threshold was set from the first FoundationPose failure per session. This threshold is used only to stratify the evaluation and does not affect VICON tracking or optimization. This leaves 24,285 occluded and 22,937 visible frames.

Accuracy is measured as the translation error $e_{T,k} = 10^3\|\mathbf{p}_k - \mathbf{p}^*_k\|_2$ (mm) and the symmetry-aware rotation error with respect to the GT pose $(\mathbf{R}^*_k, \mathbf{p}^*_k)$,

$$e_{R,k} = \frac{180}{\pi} \min_{\mathbf{S}\in\mathcal{S}_O} \cos^{-1} \frac{\mathrm{tr}[(\mathbf{R}^*_k\mathbf{S})^T\mathbf{R}_k] - 1}{2}, \tag{13}$$

where the argument of $\cos^{-1}$ is clipped to $[-1,1]$ and $\mathcal{S}_O$ holds the rotations that leave the object mesh unchanged: 90° steps about the mesh z-axis for the block, 180° for the phone (flips excluded), and only $\{I\}$ for the rest. We report the percentage of correctly tracked frames, $\mathrm{PCF}_m(\tau;\mathcal{K})$, i.e. the fraction (in %) of frames in $\mathcal{K}$ with $e_{m,k} \le \tau$ for $m \in \{T,R\}$, and abbreviate $S_{T20} = \mathrm{PCF}_T(20\,\mathrm{mm})$ and $S_{R15} = \mathrm{PCF}_R(15^\circ)$. A frame *fails* when $e_{T,k} > 20$ mm or $e_{R,k} > 15^\circ$, since a wrong orientation spoils the contact geometry even at the right position; the tolerances are commensurate because a 15° rotation displaces a surface point on these objects by roughly 20 mm. Failed-frame rates are taken over all 47,222 frames, and Fig. 5(b) also reports the normalized area under the PCF curve up to $\tau_{\max} = 60$ mm or 90°.

C. Comparison with Baseline Trackers

Our baselines are FoundationPose [16], MegaPose [18] run sequentially on RGB-D, and ICG+ [28] with its region, depth, and ORB modalities in its YCB-Video configuration. All of them saw the same RGB-D frames and meshes, started from the first FoundationPose pose registered with a text-prompted SAM 3 mask [29], and then tracked all 40 sessions without external pose feedback or reset from the GT. None

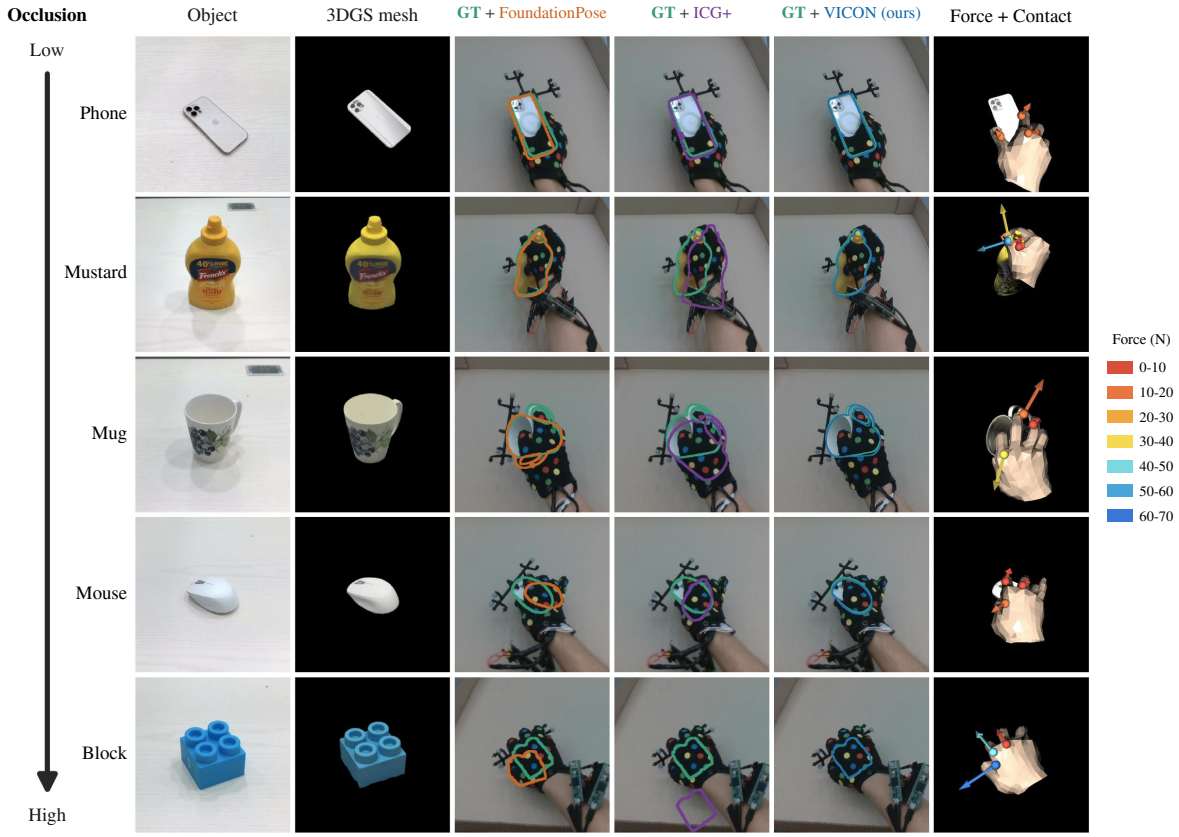


Fig. 6. Qualitative results with increasing hand occlusion from top to bottom. Columns: object, 3DGS mesh, and the input image overlaid with the ground-truth pose (green) and the pose of FoundationPose, ICG+, and VICON. Last column: rotated view of the tracked hand, visualized as a MANO [27] mesh, and the VICON object pose with the contact points and force arrows of the recognized FSRs, colored by the calibrated normal force.

TABLE III

COMPARISON WITH BASELINE TRACKERS

Method	Failed frames (%)	Success rate Visible S_{T20} (%)	Success rate Visible S_{R15} (%)	Success rate Occluded S_{T20} (%)	Success rate Occluded S_{R15} (%)	Median error Visible T (mm)	Median error Visible R (deg)	Median error Occluded T (mm)	Median error Occluded R (deg)
FoundationPose	50.9	63.6	62.5	48.8	39.6	4.0	2.7	21.8	33.9
MegaPose	62.6	79.3	64.5	26.2	15.3	4.4	5.9	45.7	82.9
ICG+	64.6	52.3	57.9	25.6	27.3	6.9	9.5	61.1	34.3
VICON (ours)	**2.5**	**100.0**	**99.6**	**97.1**	**98.4**	**0.9**	**1.7**	**3.9**	**3.0**

Failed frames: $T > 20$ mm or $R > 15°$ over all 47,222 frames (Sec. V-B). Visible/occluded: $v_d^{\mathrm{GT}} \geq 0.55$ / < 0.55 based on the threshold set from the first FoundationPose failure per session (Sec. V-B). S_{T20}/S_{R15}: PCF (Sec. V-B) at 20 mm/15°. R: symmetry-aware rotation error (13). Bold: best per column.

TABLE IV

COMPONENT ABLATIONS

Variant	Failed frames (%)	Success rate Visible S_{T20} (%)	Success rate Visible S_{R15} (%)	Success rate Occluded S_{T20} (%)	Success rate Occluded S_{R15} (%)	Median error Visible T (mm)	Median error Visible R (deg)	Median error Occluded T (mm)	Median error Occluded R (deg)
w/o vis. weighting	28.9	93.7	84.8	74.1	63.3	3.0	1.9	5.9	6.3
w/o hand motion	18.1	99.9	97.5	73.7	78.4	1.0	1.8	5.2	5.2
w/o contact timing	23.9	99.7	92.9	85.4	65.8	1.4	1.9	7.6	9.4
VICON (full)	**2.5**	**100.0**	**99.6**	**97.1**	**98.4**	**0.9**	**1.7**	**3.9**	**3.0**

Each row removes one component from the final refinement, with the contact states of Sec. IV-C held fixed. Columns as in Table III. Bold: best per column.

of the baselines was given FSR or glove measurements, and Table III compares the proposed VICON and baselines.

VICON reduces the failed-frame rate to 2.5%, compared with 50.9–64.6% for the baselines (Table III), with occluded median errors of 3.9 mm and 3.0° (Fig. 5). FoundationPose fails on 38.8% of the visible frames; all but two frames follow an occluded failure without recovery, so the errors persist once visibility returns. Sequential tracking refines from the previous frame without re-checking the estimate against the observation, whereas VICON down-weights degraded poses, carries the object through the grasp interval with the hand-motion prior, and re-measures from there. VICON has the lowest occluded median translation error in 36 of the 40 sessions, with 23 sessions containing no failures, and the per-object occluded medians range from 2.3 to 5.1 mm. Fig. 6 illustrates tracking and contact annotations under increasing hand occlusion, where FoundationPose and ICG+ deviate from the GT as the hand covers the object while VICON remains close.

D. Component Ablations

Table IV ablates the second (final) refinement on the same frames, with the hand poses and the contact states of Sec. IV-C fixed, so the variants differ only in how the refinement

uses the visual and contact evidence. Each variant removes one component and retains the other two.

Visibility weighting. The acceptance gates, hand mask, and outlier checks are removed and the observation deviations fixed to their $v_d = 0.5$ values, so unreliable FoundationPose candidates enter the graph at that fixed weight.

Hand-motion prior. We disable hand-motion initialization, propagation, the contact-gated prior, the grasp links ($\mathcal{A}$, $\mathcal{G}$), and the synthetic poses, leaving occluded frames dependent on visual evidence that is itself degraded. The occluded-frame medians include stationary periods and therefore do not isolate the contribution during transport.

Tactile contact timing. The contact gate is disabled, so the hand-motion prior is applied to every usable hand increment, including intervals in which the hand moves without the object.

Removing any component increases the failed-frame rate from 2.5% to 18.1–28.9%. Paired comparisons of the per-session occluded median translation errors favor full VICON over each ablation (40 sessions; two-sided sign tests, Holm-corrected $p < 0.0005$).

E. Dataset and Contact-Force Annotations

The current dataset consists of all 71 sessions (including 40 motion-capture sessions): 81,307 RGB-D frames, meshes, hand skeleton trajectories and refined object trajectories, and the contact records of Sec. IV-E. The calibrated normal forces have a median of 10.9 N and a 90th percentile of 27.0 N; 83% of them fall between 0 and 20 N, the range for which the calibration error is reported in Sec. IV-E.

VI. Conclusion

We present VICON, a visual-inertial-contact based hand-object tracking framework that is robust to occlusion and enables scalable collection of contact-rich demonstrations. Using a single RGB-D camera and the VICON-glove, our system synchronously captures hand-object motion, contact points, and contact forces. Furthermore, our offline refinement recovers robust object trajectories even under severe occlusion by weighting visual estimates by visibility and admitting hand motion constraints only under measured contact. Using our framework, the failed-frame rate falls from 50.9% (FoundationPose) to 2.5%. The current dataset covers single-hand sequences with five rigid objects. Future work will add subjects, object classes, and faster motions, while relaxing the fixed-grasp prior to support full in-hand manipulation, and evaluate these synchronized kinematic and tactile records in policy learning.

References

[1] Y. Lee, W. Do, H. Yoon, J. Heo, W. Lee, and D. Lee, "Visual-inertial hand motion tracking with robustness against occlusion, interference, and contact," *Sci. Robot.*, vol. 6, no. 58, p. eabe1315, 2021.

[2] R. Hoque, P. Huang, D. J. Yoon, M. Sivapurapu, and J. Zhang, "EgoDex: Learning dexterous manipulation from large-scale egocentric video," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2026.

[3] T. Tao, M. K. Srirama, J. J. Liu, K. Shaw, and D. Pathak, "DexWild: Dexterous human interactions for in-the-wild robot policies," in *Proc. Robot.: Sci. Syst. (RSS)*, 2025.

[4] C. Chen, Z. Yu, H. Choi, M. Cutkosky, and J. Bohg, "DexForce: Extracting force-informed actions from kinesthetic demonstrations for dexterous manipulation," *IEEE Robot. Autom. Lett.*, vol. 10, no. 6, pp. 6416–6423, 2025.

[5] J. Yu *et al.*, "ForceVLA: Enhancing VLA models with a force-aware MoE for contact-rich manipulation," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 38, 2025, pp. 103 609–103 639.

[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," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 361–378.

[7] Z. Fan *et al.*, "ARCTIC: A dataset for dexterous bimanual hand-object manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 12 943–12 954.

[8] 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.

[9] Y. R. Song *et al.*, "OpenTouch: Bringing full-hand touch to real-world interaction," *arXiv preprint arXiv:2512.16842*, 2025.

[10] S. Ni *et al.*, "TactiDex: A real-world tactile-guided benchmark for human-like dexterous manipulation," *arXiv preprint arXiv:2607.09190*, 2026.

[11] J. Park, Y. Oh, G. Moon, H. Choi, and K. M. Lee, "HandOccNet: Occlusion-robust 3D hand mesh estimation network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 1486–1495.

[12] 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 *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 294–310.

[13] H. Qi, C. Zhao, M. Salzmann, and A. Mathis, "HOISDF: Constraining 3D hand-object pose estimation with global signed distance fields," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 10 392–10 402.

[14] Z. Fan *et al.*, "HOLD: Category-agnostic 3D reconstruction of interacting hands and objects from video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 494–504.

[15] J. Shin *et al.*, "ComPose: When to trust hands for object pose tracking," *arXiv preprint arXiv:2605.23523*, 2026.

[16] B. Wen, W. Yang, J. Kautz, and S. Birchfield, "FoundationPose: Unified 6D pose estimation and tracking of novel objects," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 17 868–17 879.

[17] F. Gonzalez, F. Gosselin, and W. Bachta, "Analysis of hand contact areas and interaction capabilities during manipulation and exploration," *IEEE Trans. Haptics*, vol. 7, no. 4, pp. 415–429, 2014.

[18] Y. Labbé *et al.*, "MegaPose: 6D pose estimation of novel objects via render & compare," in *Proc. Conf. Robot Learn. (CoRL)*, ser. Proc. Mach. Learn. Res., vol. 205, 2023, pp. 715–725.

[19] H. Choi *et al.*, "Robust hand tracking from visual-inertial fusion," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2026.

[20] Y. Li, J. Zhang, W. Xu, T. Tang, and C. Lu, "FSGlove: An inertial-based hand tracking system with shape-aware calibration," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2025, pp. 18 304–18 311.

[21] B. Wen, S. Dewan, and S. Birchfield, "Fast-FoundationStereo: Real-time zero-shot stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2026, pp. 7513–7524.

[22] Interlink Electronics, "FSR 400 series data sheet," https://www.interlinkelectronics.com/downloads/datasheets/fsr-400-series-datasheet.pdf, P/N PDS-10004-C, accessed Sep. 12, 2026.

[23] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 4104–4113.

[24] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3D Gaussian splatting for real-time radiance field rendering," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139:1–139:14, 2023.

[25] A. Guédon and V. Lepetit, "SuGaR: Surface-aligned Gaussian splatting for efficient 3D mesh reconstruction and high-quality mesh rendering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 5354–5363.

[26] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar, "The YCB object and model set: Towards common benchmarks for manipulation research," in *Proc. Int. Conf. Adv. Robot. (ICAR)*, 2015, pp. 510–517.

[27] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: Modeling and capturing hands and bodies together," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 245:1–245:17, 2017.

[28] M. Stoiber, M. Elsayed, A. E. Reichert, F. Steidle, D. Lee, and R. Triebel, "Fusing visual appearance and geometry for multi-modality 6DoF object tracking," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2023, pp. 1170–1177.

[29] N. Carion *et al.*, "SAM 3: Segment anything with concepts," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2026.