Visuo-Acoustic Hand Pose and Contact Estimation

Yuemin Mao [1*] **Uksang Yoo**[1*] **Yunchao Yao**[1] **Shahram Najam Syed**[1]

Luca Bondi[2] **Jonathan Francis** [1,2] **Jean Oh**[1] **Jeffrey Ichnowski**[1]

[*]Equal contribution.

[1]Robotics Institute, Carnegie Mellon University, Pittsburgh, USA

[2]Bosch Center for Artificial Intelligence, Pittsburgh, USA

Abstract

Accurately estimating hand pose and hand-object contact events is essential for robot data-collection, immersive virtual environments, and biomechanical analysis, yet remains challenging due to visual occlusion, subtle contact cues, limitations in vision-only sensing, and the lack of accessible and flexible tactile sensing. We therefore introduce **VibeMesh**, a novel wearable system that fuses vision with active acoustic sensing for dense, per-vertex hand contact and pose estimation. VibeMesh integrates a bone-conduction speaker and sparse piezoelectric microphones, distributed on a human hand, emitting structured acoustic signals and capturing their propagation to infer changes induced by contact. To interpret these cross-modal signals, we propose a graph-based attention network that processes synchronized audio spectra and RGB-D-derived hand meshes to predict contact with high spatial resolution. We contribute: (i) a lightweight, non-intrusive visuo-acoustic sensing platform; (ii) a cross-modal graph network for joint pose and contact inference; (iii) a dataset of synchronized RGB-D, acoustic, and ground-truth contact annotations across diverse manipulation scenarios; and (iv) empirical results showing that VibeMesh outperforms vision-only baselines in accuracy and robustness, particularly in occluded or static-contact settings.

1 Introduction

Accurately estimating human hand pose *and* contact is critical for robot teleoperation [30, 59, 9, 55], virtual reality [52, 1], and biomechanical analysis [10, 35, 31]. In all of these settings, knowing *when* and *where* the hand touches the environment—together with its configuration—enables reasoning about task phases, distinguishing exploration from manipulation, and inferring force dynamics. Unfortunately, real-world contact sensing is hard: occlusions, limited sensor viewpoints, and subtle touch events routinely confound purely visual approaches.

Vision-based methods typically estimate contact indirectly, pairing RGB or depth observations with strong priors on object geometry and canonical hand poses [14, 4, 54, 41]. Model-based fitting can help [47, 48, 20], but still fails under poor lighting and suffers from ambiguities due to occlusion. Direct tactile solutions, for example, capacitive and piezoelectric gloves [24, 18] or full-body suits [15], offer better signal fidelity at the cost of bulk, expense, and limited practicality. Meanwhile, promising advances in wearable acoustics [57] and cross-modal learning [43] have yet to be exploited for dense hand-object contact estimation.

We propose bridging this gap with a *visuo-acoustic* sensing approach that delivers joint pose- and contact-estimation. A lightweight bone-conduction speaker, mounted on the wrist, emits a signal

Preprint. Under review.

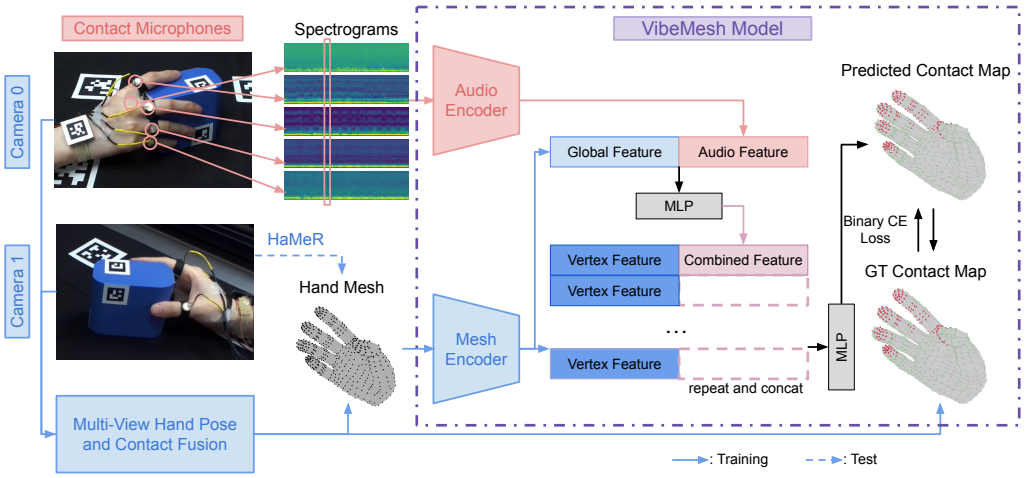


Figure 1: **Overview of VibeMesh.** Our visuo-acoustic contact estimation architecture predicts per-vertex contact by integrating audio embeddings with hand-mesh features. During training, it leverages audio with hand meshes and contact annotations jointly reconstructed from two synchronized RGB-D camera streams; at inference, it operates on audio and a hand mesh reconstructed from a single, partially occluded, RGB view, demonstrating robustness under visual ambiguity.

that consists of a wide range of acoustic frequencies (a process called 'broadband probing'), whose propagation behavior changes whenever the hand changes configuration or touches an object. A sparse array of piezoelectric contact microphones on the fingers records these shifts, providing rich, self-generated cues that remain informative even when the hand is static or visually occluded. To interpret them, we propose **VibeMesh**: a cross-modal graph-attention network that fuses spectral audio features with a mesh-based MANO [33] hand representation to predict per-vertex contact labels from synchronized audio and RGB input (Fig. 1).

This paper contributes: (*i*) a wearable visuo-acoustic platform for contact-aware hand tracking; (*ii*) VibeMesh, a cross-modal graph-attention architecture that fuses acoustic and visual cues for dense contact prediction; (*iii*) a dataset of time-aligned RGB-D, audio, and ground-truth contact annotations across diverse grasps; and (*iv*) thorough evaluation across users and objects demonstrating improved accuracy and robustness under occlusions, diverse object properties, and static-contact scenarios compared against state-of-the-art baselines.

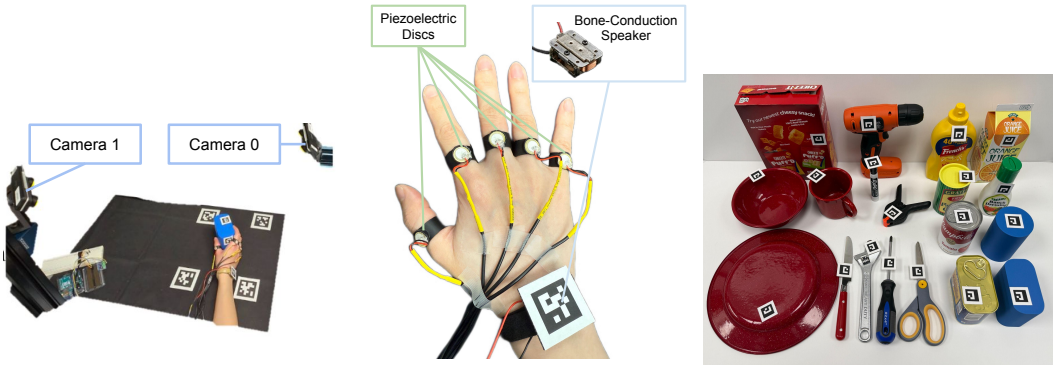
2 Related Works

2.1 Visual Hand Pose and Contact Estimation

Researchers have extensively studied vision-based hand pose and contact estimation motivated by its application to VR/AR/XR [52, 1, 25, 5], robotics [30, 59, 9, 55, 27], and kinesiology [2, 46]. The ubiquity of cameras in wearable devices and everyday environments makes vision a convenient and accessible sensing modality [58].

Classical approaches primarily relied on template matching and silhouette-based techniques [12, 34]. More recent methods leverage large-scale synthetic [22, 7] or real-world datasets [53, 39, 26, 8, 16, 61] and expressive models to regress 2D keypoints or full 3D meshes from monocular RGB inputs [60, 14]. Transformer-based architectures [19, 17, 51, 29] and physics-informed pipelines [47, 45] have further improved pose estimation accuracy and robustness.

However, vision-only methods remain fundamentally limited by self-occlusion and inherent visual ambiguities, particularly during hand-object interactions [13]. To address these limitations, VibeMesh incorporates audio as an additional sensing modality. By exploiting the rich temporal and spectral structure of contact-induced sounds, which reflect changes in the lumped acoustic properties of the hand-object system, VibeMesh learns to disambiguate visually similar hand poses and to localize contact more precisely than vision-only approaches.



(a) Data collection workspace. (b) Wearable microphone rings. (c) Data collection objects.

Figure 2: **VibeMesh data collection setup.** (a) Our multi-view capture station with two calibrated ZED Mini RGB-D cameras positioned at complementary angles to minimize occlusion during hand-object interaction. (b) Custom-designed piezoelectric microphone rings worn on each finger, with the bone-conduction speaker mounted on the wrist. Each ring contains a 10 mm piezoelectric disc sensor connected to a Maono USB sound card. (c) The diverse set of 19 objects used in our experiments includes items from YCB and HOPE datasets with varied geometries, materials, and affordances to evaluate across different manipulation scenarios.

2.2 Multi-Modal Human Pose Tracking

To address the inherent limitations of vision-only systems, recent work has explored incorporating additional sensing modalities such as tactile [28, 42, 11], EMG [37, 56, 36, 23], and audio [21, 50, 49, 38]. These modalities provide complementary signals that capture muscle activity, contact forces, and acoustic transients associated with physical interactions, enabling robust pose and contact estimation in the presence of occlusion or visual ambiguity.

Tactile and stretchable sensors offer direct measurements of contact and deformation but often require expensive, form-fitting gloves or embedded skins that can restrict natural user motion [3]. EMG-based approaches infer pose from muscle activation, yet are sensitive to electrode placement and prone to signal drift or degradation due to user fatigue [36, 40]. Additionally, tactile sensor-embedded gloves generally suffer from low spatial resolution, often segmenting the hand into a few contact patch regions [37]. In contrast, audio sensing offers an indirect but non-intrusive alternative [57], with sound carrying rich temporal and spectral cues indicative of contact timing, material properties, and dynamic interactions—even when visual signals are occluded. By leveraging synchronized audio-visual signals, VibeMesh learns to disambiguate occluded hand poses and accurately localizes contact events, while minimizing reliance on wearable or invasive instrumentation.

3 Methods

VibeMesh consists of three stages: data collection, vision-based ground-truth contact labeling, and model training. In the data collection stage, we capture multi-model recordings by synchronizing two RGB-D cameras with a set of wearable acoustic sensors. Next, we merge the video streams to produce ground-truth 3D hand poses and contact annotations. For model training, we learn a network that fuses mesh features from a graph-attention encoder with CNN-extracted audio embeddings, optimizing contact prediction via binary cross-entropy against the ground-truth contact labels.

3.1 Visuo-Acoustic Hand-Object Interaction Data Collection

We develop a system that integrates wearable acoustic sensors with RGB-D cameras to capture multi-model data for hand-object interactions. We collect a dataset from multiple users performing a wide range of grasping conditions.

Wearable Microphones and Speaker For active acoustic sensing, we use an Adafruit bone-conduction speaker to generate white noise as the signal, paired with five contact microphones

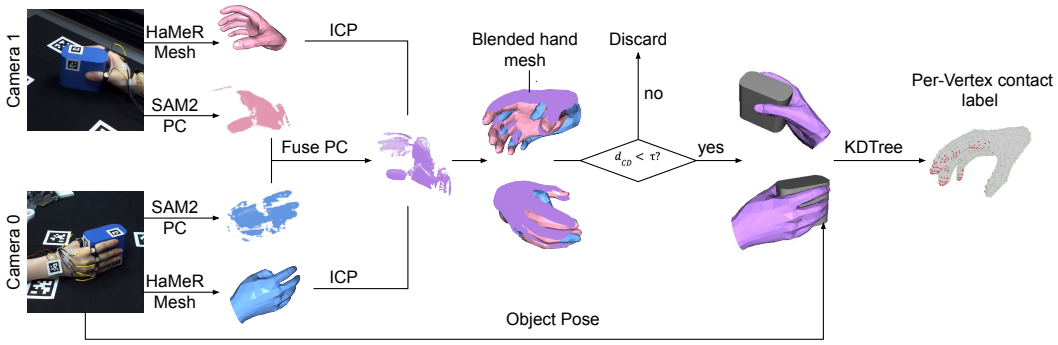


Figure 3: **VibeMesh ground-truth hand pose and contact annotation pipeline.** Our multi-view dataset collection approach integrates data from two RGB-D cameras to generate accurate hand meshes and contact annotations. Starting with RGB input from both cameras, we use SAM2 [32] for hand segmentation to extract point clouds (PC), while HaMeR [29] generates initial MANO-based hand meshes. These components are fused through calibrated camera extrinsics, with ICP registration aligning the meshes to the combined point cloud. The blended mesh is validated against a chamfer distance threshold (d_{CD}), discarding frames with poor alignment. For valid frames, we track object pose using ArUco markers with ICP refinement, then we compute proximity between hand vertices and the object surface, generating per-vertex binary contact labels with a 5-mm threshold. This pipeline creates high-quality ground-truth data for training our visuo-acoustic model even in challenging contact-rich scenarios.

(10 mm piezoelectric discs) for signal acquisition at a sampling rate of 44.1 kHz. Each contact microphone is interfaced with a Maono USB sound card and mounted onto a 3D printed PLA ring, sized appropriately for specific user fit. During data collection, participants wear the rings on their fingers, with the speaker secured to their wrist using an elastic band (Fig. 2b).

RGB-D Cameras To obtain high-quality ground-truth labels for hand pose and contact, we use two ZED Mini stereo RGB-D cameras positioned at complementary viewing angles (Fig. 2a) to reduce occlusions. Both cameras record synchronized data at a resolution of 1080p and a frame rate of 30 fps, and operate with Zed AI disparity estimation to enhance depth map resolution and accuracy.

Data Collection Procedure Five right-handed participants (3 male and 2 female) with varied hand and finger dimensions took part in data collection under IRB approval. We selected 19 objects with high-quality meshes varying in geometry, mass, and material (Fig. 2c): 14 from the YCB dataset [6], 3 from the HOPE dataset [44], and 2 PLA-printed replicas matching 2 YCB objects. Each participant completed five 60-second sessions per object, during which they repeatedly grasped and released the item and were instructed to creatively explore different grasp poses to ensure a diversity of hand poses and contact conditions.

3.2 Vision-Based Hand Pose and Contact Estimation

We take a vision-based approach to obtain ground-truth contact labels (Fig. 3), addressing occlusion by fusing hand reconstructions from both camera views. For object tracking, we use ArUco markers with Iterative Closest Point (ICP) refinement to balance accuracy and efficiency.

Multi-view Hand Pose Fusion To reconstruct the hand pose, we integerate RGB-D from 2 complementary views with learned mesh priors. For each 60-s recording, we segment 3 keyframes per camera using SAM2 [32], and propagate these masks to extract hand point clouds across time. In parallel, HaMeR [29] estimates anatomically plausible hand meshes with 778 vertices using MANO [33], which enforces consistent topology and incorporates learned kinematic priors. We first fuse the two hand point clouds using calibrated camera extrinsics and align each mesh to the fused point cloud via ICP. Then we merge the two ICP-aligned meshes using As-Rigid-As-Possible (ARAP) deformation, guided by the local geometry agreement. If the blended mesh fails to align with the fused point cloud, as measured by a chamfer distance threshold, we discard the corresponding

frames to enhance label reliability. To ensure robust, stable tracking, we also mount an ArUco marker on the user's wrist and perform an additional ICP alignment to that marker's pose, constraining the hand mesh to a reasonable global location. This joint fusion gives an accurate, temporally coherent 3D reconstruction of the hand.

Object Tracking We attach an ArUco marker to each object and detect its 6-DoF pose relative to the camera in every frame. Using this initial estimation, we align the object mesh accordingly. To improve accuracy, the pose is further refined via ICP registration between the mesh and the depth data. This refinement yields accurate, frame-wise tracking of the object's position and orientation.

Contact Estimation With hand and object meshes registered in a shared coordinate frame, we assign a binary contact label to each of hand mesh vertices based on proximity. A vertex is marked in contact if its nearest point on the object lies within a 5-mm threshold. This approach yields dense, frame-level contact annotations that are geometrically consistent and robust to occlusion.

3.3 VibeMesh Model

The VibeMesh architecture (Figure 1) consists of three main components: an audio encoder, a mesh encoder, and a cross-modal fusion module for contact prediction.

Audio Encoder We process the continuous multi-channel acoustic signals to isolate and enhance contact-relevant features through the following steps: (*i*) *reference subtraction*, where a baseline acoustic profile, captured during the first 5 ms of each recording when the hand is free-floating without contact, is subtracted from subsequent frames to isolate contact-induced changes; (*ii*) *temporal alignment*, where the waveform is segmented into 35-ms windows corresponding to each video frame (30 fps) with shifting windows to maintain synchronization between modalities; (*iii*) *spectrogram extraction*, using a short-time Fourier transform (1024-point FFT, 512-point hop length) to yield time-frequency representations for all five microphone channels; and (*iv*) *normalization*, applying frequency-bin normalization followed by per-channel standardization to ensure consistent feature representation and mitigate variations in microphone sensitivity introduced by various factors such as manufacturing variance.

This preprocessing yields aligned spectrograms that capture the characteristic spectral changes when the hand contacts objects. To extract discriminative features from these spectrograms, we use a pretrained VGG backbone that we finetune. The network processes each microphone channel independently before channel-wise feature fusion through self-attention, allowing the model to adaptively weight signals based on their relevance to contact events. The final audio embedding vector $\mathbf{z}_{\text{audio}} \in \mathbb{R}^{256}$ encodes the complex acoustic patterns resulting from both hand pose configuration and object contact interactions, providing complementary information to the visual modality.

Mesh Encoder The mesh encoder processes the hand geometry using a hierarchical graph neural network that preserves the mesh's topological structure. Given a hand mesh with $N = 778$ vertices and an adjacency matrix derived from mesh connectivity, we apply a series of graph convolutional operations:

$$\mathbf{H}^{(1)} = \text{ReLU}(\text{GCN}(\mathbf{X}, \mathbf{A})) \tag{1}$$

$$\mathbf{H}^{(2)} = \text{ReLU}(\text{GCN}(\mathbf{H}^{(1)}, \mathbf{A})) \tag{2}$$

$$\mathbf{H}^{(3)} = \text{GAT}(\mathbf{H}^{(2)}, \mathbf{A}), \tag{3}$$

where $\mathbf{X} \in \mathbb{R}^{N\times 3}$ represents vertex coordinates, $\mathbf{A}$ is the adjacency matrix, GCN denotes graph-convolutional layers, and GAT represents a graph-attention layer with 4 attention heads. The resulting node embeddings $\mathbf{H}^{(3)} \in \mathbb{R}^{N\times 256}$ capture local geometric features for each vertex. In our notation, $\mathbf{H}^{(l)}$ represents the feature matrix for all vertices at the l-th layer of the graph neural network, while $\mathbf{h}_i^{(l)}$ denotes the feature vector for vertex i at layer l. We compute a global mesh representation using a global attention pooling layer:

$$\mathbf{z}_{\text{mesh}} = \sum i = 1^N \alpha_i \cdot \mathbf{h}^{(3)}. \tag{4}$$

This hierarchical approach enables the model to capture both local vertex-level features and global hand configuration information. A key insight with this graph-based approach is that the hand's

kinematics and use often creates long-range dependencies among contacts, where grasping typically engages opposing surfaces and are pose-dependent. The VibeMesh architecture leverages this with progressively larger receptive fields in early layers followed by attention mechanism that can model anatomically far but functionally correlated contact regions. This information is pooled together while preserving relevant contact cues. This information is complemented by the acoustic modality to disambiguate contact conditions given the present pose of the hand.

Cross-Modal Fusion and Contact Prediction The fusion module combines audio and mesh representations to predict per-vertex contact probabilities. First, we concatenate global features from both modalities $\mathbf{z}_{\text{global}} = [\mathbf{z}_{\text{audio}}; \mathbf{z}_{\text{mesh}}]$.

This combined representation is processed through an MLP to extract cross-modal features:

$$\mathbf{z}_{\text{fused}} = f(\mathbf{z}_{\text{global}}) \tag{5}$$

We then compute per-vertex contact predictions by combining local vertex features with the global cross-modal representation:

$$\mathbf{v}_i = [\mathbf{h}_i^{(3)}; \mathbf{z}_{\text{fused}}] \tag{6}$$

$$\alpha_i = \sigma(g_a(\mathbf{v}_i)) \tag{7}$$

$$\hat{y}_i = \sigma(g_p(\mathbf{v}_i \odot \alpha_i)), \tag{8}$$

where $\mathbf{h}_i^{(3)}$ represents the feature vector for vertex i, α_i is an attention weight, and $\hat{y}_i$ is the predicted contact probability. This attention mechanism allows the model to focus on the hand's most relevant regions based on geometric and acoustic cues.

3.3.1 Training Procedure

We train our model using a weighted binary cross-entropy loss to address the significant class imbalance inherent in contact estimation, where contact vertices typically constitute only 5–10 % of the total vertices:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[w_1 y_i \log(\hat{y}_i) + w_0(1-y_i)\log(1-\hat{y}_i)], \tag{9}$$

where $y_i \in \{0, 1\}$ is the ground-truth contact label for vertex i, and w_1 and w_0 are class weights for the positive (contact) and negative (non-contact) classes, respectively. We compute these weights inversely proportional to class frequencies in each training batch:

$$w_1 = \frac{N}{2 \cdot \sum_{i=1}^{N} y_i}, \quad w_0 = \frac{N}{2 \cdot \sum_{i=1}^{N}(1-y_i)}. \tag{10}$$

This weighting scheme penalizes false negatives more heavily than false positives, encouraging the model to correctly identify the sparse contact regions despite their underrepresentation in the training data.

3.4 Implementation Details

VibeMesh's mesh encoder has a hierarchical graph neural network with two GCN layers (64 and 128 channels) followed by a 4-headed graph attention layer (256 channels). We use the MANO [33] hand mesh topology to define edge connectivity for message passing operations. For training, we used the Adam optimizer with an initial learning rate of 0.001, batch size of 32 with gradient accumulation for effective batch sizes of 512, and a reduce-on-plateau scheduler with a factor of 0.5 and patience of 5 epochs. To mitigate overfitting, we utilize dropouts ($p = 0.2$–0.3) throughout the network. The models were trained for 20 epochs on NVIDIA RTX 4090 GPUs. Each model took around 15 hours to train.

4 Experiments

We conducted an evaluation of VibeMesh to assess its effectiveness in hand-object contact estimation across diverse scenarios. Our experiments were designed to validate key aspects: (1) the accuracy

Table 1: Performance comparison across different input modalities. We report average F1 Score (↑) and average Chamfer Distance (↓), along with performance on unseen objects and subjects.

Method	Modality	Avg F1 (↑)	Avg Chamfer (↓)	Unseen Obj F1 (↑)	Unseen Subj F1 (↑)	Unseen Obj Chamfer (↓)	Unseen Subj Chamfer (↓)
Hold [14] (Baseline)	RGB	0.1171 ± 0.0974	31.33 ± 27.36 mm	0.1431 ± 0.1026	0.1325 ± 0.0889	46.55 ± 13.45 mm	19.92 ± 5.489 mm
VibeMesh (Proposed)	RGB + Audio	**0.327** ± 0.122	**5.487** ± 1.612 mm	**0.288** ± 0.111	**0.302** ± 0.116	**6.042** ± 1.755 mm	**6.828** ± 1.704 mm

of the proposed visuo-acoustic approach compared to a vision-only baseline, (2) the robustness of the system under challenging conditions like occlusion and complex object geometry (3) the generalization capabilities across unseen objects and users.

We first describe our evaluation metrics to test both contact classification-based accuracy measures and geometric accuracy. We then present quantitative results comparing VibeMesh to state-of-the-art vision-only baselines, followed by ablation studies that isolate the contribution of each modality and architectural component. Finally, we showcase qualitative results highlighting specific scenarios where our approach excels, particularly in cases where vision alone struggles to accurately determine contact regions.

4.1 Evaluation Metrics

We evaluate contact estimation with two complementary metrics: label accuracy and geometric precision. Both provide insight into how well VibeMesh may work in comparison to state-of-the-art baselines. For label accuracy, we use the F1 score, defined as:

$$F_1 = \frac{2 \cdot P \cdot R}{P + R}, \tag{11}$$

where P, precision, represents the fraction of predicted contacts that are true: $P = \frac{|\mathcal{C}_{\text{pred}} \cap \mathcal{C}_{\text{true}}|}{|\mathcal{C}_{\text{pred}}|}$, and R, recall, quantifies the fraction of true contacts that are predicted: $R = \frac{|\mathcal{C}_{\text{pred}} \cap \mathcal{C}_{\text{true}}|}{|\mathcal{C}_{\text{true}}|}$. This metric effectively penalizes both false positives and false negatives in contact estimation, and a higher F1 score indicates better performance.

For geometric precision, we compute the chamfer distance (d_{CD}) between the set of predicted contact vertices $\mathcal{V}_{\text{pred}}$ and the set of ground-truth contact vertices $\mathcal{V}_{\text{true}}$:

$$d_{CD}(\mathcal{V}_{\text{pred}}, \mathcal{V}_{\text{true}}) = \frac{1}{2|\mathcal{V}_{\text{pred}}|} \sum_{x \in \mathcal{V}_{\text{pred}}} \min_{y \in \mathcal{V}_{\text{true}}} \|x - y\| + \frac{1}{2|\mathcal{V}_{\text{true}}|} \sum_{y \in \mathcal{V}_{\text{true}}} \min_{x \in \mathcal{V}_{\text{pred}}} \|x - y\|. \tag{12}$$

This metric averages the squared nearest-neighbor distances in both directions between the two sets, so a low chamfer distance indicates that predicted contact vertices lie very close in $\mathbb{R}^3$ to the true contacts. Together, the F1 score and chamfer distance ensure our evaluation captures both correct classification of contact regions and their localization with high spatial fidelity.

4.2 Results

Our experiments demonstrate that VibeMesh noticeably outperforms vision-only and audio-only approaches across various scenarios. Table 2 shows that VibeMesh achieves an average F1 score of 0.327, representing a 179.4% improvement over the state-of-the-art vision-only Hold [14] baseline (F1 score of 0.117). The substantial performance gap is evident in the chamfer distance metric, where VibeMesh (5.487 mm) achieves an 82.5% reduction compared to the vision-only baseline (31.33 mm). This supports the complementary nature of the visuo-acoustic approach, where each modality contributes unique information about contact conditions, which we study further in ablation studies in the later sections.

When analyzing performance on the challenging case of unseen objects, VibeMesh demonstrates robust generalization capabilities with an F1 score of 0.288, significantly outperforming the vision-only baseline (0.143) by 101.4%. This suggests that our cross-modal approach captures generalizable features of hand-object interactions that transfer effectively to novel objects of different geometries and materials.

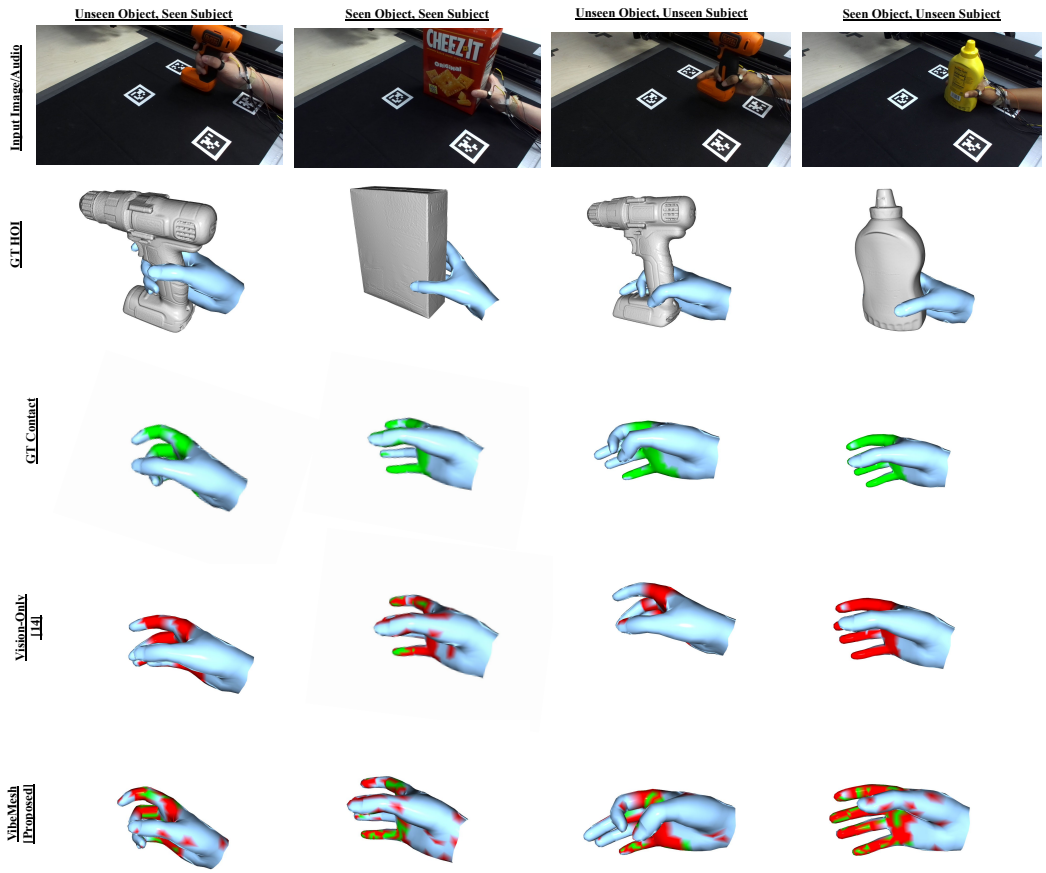


Figure 4: **Qualitative Results**. Each column represents a different test condition: unseen object with seen subject, seen object with seen subject, unseen object with unseen subject, and seen object with unseen subject. The rows show: (1) input RGB images to the models with the subjects interacting with objects while wearing the active acoustic sensing platform (2) ground-truth hand-object interaction (GT HOI) visualizations showing 3D hand meshes and object models (3) ground-truth contact labels with contact vertices highlighted in green (4) contact prediction results from the vision-only baseline [14], showing substantial false negatives (missing contacts) and false positives (incorrect contact regions) in red (5) VibeMesh predictions, demonstrating noticeable improvements in contact localization across all generalization scenarios. We note that VibeMesh largely identifies contact points even in challenging cases with partial occlusion and novel objects/subjects.

Similarly, the chamfer distance on unseen objects (6.042 mm) shows an 87.0% improvement over the vision-only approach (46.55 mm), indicating that our method accurately localizes contact points even on previously unseen geometries. Cross-user evaluation further highlights VibeMesh's generalization abilities across different hand sizes and interaction styles, maintaining an average F1 score of 0.302. This represents a 7.6% decrease from the overall average, indicating that the learned visuo-acoustic features capture fundamental patterns of hand-object contact that transcend individual differences in hand geometry and manipulation behavior.

The cross-user chamfer distance (6.828 mm) shows a 65.7% improvement over the vision-only baseline (19.92 mm), demonstrating that our approach maintains spatial precision even when applied to previously unseen hand geometries. Qualitative results in Figure 4 illustrate several challenging scenarios where VibeMesh excels. Under partial occlusion of object geometry and hand, where the vision-only baseline struggles to accurately determine contact regions, VibeMesh 's visuo-acoustic approach correctly identifies contact regions by leveraging the acoustic signals that propagate through the hand.

Table 2: Performance comparison across different input modalities. We report average F1 Score (↑) and average Chamfer Distance (↓), along with performance on unseen objects and subjects.

Method	Modality	Avg F1 (↑)	Avg Chamfer (↓)	Unseen Obj F1 (↑)	Unseen Subj F1 (↑)	Unseen Obj Chamfer (↓)	Unseen Subj Chamfer (↓)
w/o Audio	RGB	0.096 ± 0.071	9.853 ± 2.140	0.078 ± 0.062	0.082 ± 0.060	10.215 ± 2.356	9.934 ± 2.218
w/o Vision	Audio	0.218 ± 0.103	7.642 ± 1.872	0.186 ± 0.092	0.195 ± 0.098	8.124 ± 2.065	7.985 ± 1.943
w/o Fusion Module	RGB + Audio	0.287 ± 0.115	6.319 ± 1.743	0.251 ± 0.108	0.265 ± 0.112	$6.875\,\text{mm} \pm 1.826$	$6.742 \pm 1.798\,\text{mm}$
VibeMesh	RGB + Audio	$\mathbf{0.327} \pm 0.122$	$\mathbf{5.487} \pm 1.612\,\text{mm}$	$\mathbf{0.288} \pm 0.111$	$\mathbf{0.302} \pm 0.116$	$\mathbf{6.042} \pm 1.755\,\text{mm}$	$\mathbf{6.828} \pm 1.704\,\text{mm}$

4.3 Baselines and Ablation

We conduct an ablation study to examine the contribution of each component in the VibeMesh architecture, with results presented in Table 2. Removing the audio modality ("w/o Audio") results in a 70.6% decrease in F1 score and an 80.0% increase in chamfer distance, confirming that acoustic signals provide critical information for contact estimation that vision alone cannot capture. This degradation is most pronounced during contact-rich cases, where visual cues become less reliable.

Conversely, the "w/o Vision" variant—which relies solely on acoustic features—shows a 33.3% reduction in F1 score compared to the full model. While acoustic sensing excels at detecting transient events and materials properties, it lacks the spatial precision that visual information provides, particularly for localizing contacts on the hand mesh. This underscores the complementary nature of the two modalities.

The "w/o Fusion Module" variant, which processes audio and visual features independently before simple concatenation, shows a 12.2% decrease in F1 score. This highlights the importance of our cross-modal attention mechanism for integrating information from both modalities. The attention-based fusion learns modality-specific reliability, focusing on acoustic cues when vision is unreliable and leveraging visual precision when available.

5 Conclusion

We present VibeMesh, a novel visuo-acoustic approach for hand pose and contact estimation that integrates lightweight wearable acoustic sensors with visual observation. Our results demonstrate that this multi-modal system significantly outperforms vision-only approaches, particularly in challenging scenarios involving occlusions and static contacts. By fusing complementary information from acoustic propagation patterns and visual hand reconstruction, VibeMesh achieves more robust and accurate contact estimation across diverse objects and users.

VibeMesh's key insight to the task is that hand-object contacts modify the acoustic transmission properties of the hand in ways that can be measured and interpreted, even when vision may face contact condition ambiguities and fail to capture these interactions. VibeMesh's graph-based attention network effectively integrates these cross-modal signals, learning to leverage the strengths of each modality while compensating for their individual limitations. The result is a contact estimation system that maintains high performance across varied conditions.

6 Limitations

While VibeMesh demonstrates consistent improvements over vision-only approaches, it has limitations. First, although the piezoelectric contact microphones mechanically reject ambient noise by attending to only solid contact-based sound transmission, we did not include results on VibeMesh's robustness to exceptionally loud environments. However, we note that although common ambient sounds such as verbal conversations were present during data collection, the microphone signals were largely unaffected. Second, we relied on multi-view camera systems with fiducial markers for perceiving hand poses and hand-object interactions, which are inherently indirect methods of observing contacts. Although there are some commercially available force or contact sensorized gloves that could enable us to collect hand-object interactions directly, they are expensive and suffer from low spatial resolution. Finally, VibeMesh largely treats pose and contact estimation in a sequential manner, where we first estimate pose with a visual model and estimate contact conditioned on both the pose estimation and acoustic signals. In the future, we may improve on the approach by integrating both vision and acoustic signals to simultaneously reason about pose and contact to benefit from both modalities at each step.

References

[1] Ammar Ahmad, Cyrille Migniot, and Albert Dipanda. Hand pose estimation and tracking in real and virtual interaction: A review. *Image and Vision Computing*, 89:35–49, 2019.

[2] Giuseppe Airò Farulla, Daniele Pianu, Marco Cempini, Mario Cortese, Ludovico O Russo, Marco Indaco, Roberto Nerino, Antonio Chimienti, Calogero M Oddo, and Nicola Vitiello. Vision-based pose estimation for robot-mediated hand telerehabilitation. *Sensors*, 16(2):208, 2016.

[3] Valerio Belcamino, Alessandro Carfì, and Fulvio Mastrogiovanni. A systematic review on custom data gloves. *IEEE Transactions on Human-Machine Systems*, 2024.

[4] Samarth Brahmbhatt, Chengcheng Tang, Christopher D Twigg, Charles C Kemp, and James Hays. Contactpose: A dataset of grasps with object contact and hand pose. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 361–378. Springer, 2020.

[5] Paweł Buń, Jozef Husár, and Jakub Kaščak. Hand tracking in extended reality educational applications. In *International scientific-technical conference MANUFACTURING*, pages 317–325. Springer, 2022.

[6] Berk Calli, Arjun Singh, James Bruce, Aaron Walsman, Kurt Konolige, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. Yale-cmu-berkeley dataset for robotic manipulation research. *The International Journal of Robotics Research*, 36(3):261–268, 2017. doi: 10.1177/0278364917700714. URL `https://doi.org/10.1177/0278364917700714`.

[7] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12417–12426, 2021.

[8] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9044–9053, 2021.

[9] Xuxin Cheng, Jialong Li, Shiqi Yang, Ge Yang, and Xiaolong Wang. Open-television: Teleoperation with immersive active visual feedback. *arXiv preprint arXiv:2407.01512*, 2024.

[10] Jérôme Coupier, Samir Hamoudi, Sonia Telese-Izzi, Véronique Feipel, Marcel Rooze, and Serge Van Sint Jan. A novel method for in-vivo evaluation of finger kinematics including definition of healthy motion patterns. *Clinical Biomechanics*, 31:47–58, 2016.

[11] Roberto De Fazio, Vincenzo Mariano Mastronardi, Matteo Petruzzi, Massimo De Vittorio, and Paolo Visconti. Human–machine interaction through advanced haptic sensors: A piezoelectric sensory glove with edge machine learning for gesture and object recognition. *Future Internet*, 15(1):14, 2022.

[12] Ali Erol, George Bebis, Mircea Nicolescu, Richard D Boyle, and Xander Twombly. Vision-based hand pose estimation: A review. *Computer Vision and Image Understanding*, 108(1-2): 52–73, 2007.

[13] Zicong Fan, Adrian Spurr, Muhammed Kocabas, Siyu Tang, Michael J Black, and Otmar Hilliges. Learning to disambiguate strongly interacting hands via probabilistic per-pixel part segmentation. In *2021 International Conference on 3D Vision (3DV)*, pages 1–10. IEEE, 2021.

[14] Zicong Fan, Maria Parelli, Maria Eleni Kadoglou, Xu Chen, Muhammed Kocabas, Michael J Black, and Otmar Hilliges. Hold: Category-agnostic 3d reconstruction of interacting hands and objects from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 494–504, 2024.

[15] Yuki Fujimori, Yoshiyuki Ohmura, Tatsuya Harada, and Yasuo Kuniyoshi. Wearable motion capture suit with full-body tactile sensors. In *2009 IEEE International Conference on Robotics and Automation*, pages 3186–3193. IEEE, 2009.

[16] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3196–3206, 2020.

[17] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11090–11100, 2022.

[18] Lee J Hubble and Joseph Wang. Sensing at your fingertips: glove-based wearable chemical sensors. *Electroanalysis*, 31(3):428–436, 2019.

[19] Changlong Jiang, Yang Xiao, Cunlin Wu, Mingyang Zhang, Jinghong Zheng, Zhiguo Cao, and Joey Tianyi Zhou. A2j-transformer: Anchor-to-joint transformer network for 3d interacting hand pose estimation from a single rgb image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8846–8855, 2023.

[20] Zengsheng Kuang, Changxing Ding, and Huan Yao. Learning context with priors for 3d interacting hand-object pose estimation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 768–777, 2024.

[21] Chi-Jung Lee, Ruidong Zhang, Devansh Agarwal, Tianhong Catherine Yu, Vipin Gunda, Oliver Lopez, James Kim, Sicheng Yin, Boao Dong, Ke Li, Mose Sakashita, Francois Guimbretiere, and Cheng Zhang. Echowrist: Continuous hand pose tracking and hand-object interaction recognition using low-power active acoustic sensing on a wristband. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, page 1–21. ACM, May 2024. doi: 10.1145/3613904.3642910. URL `http://dx.doi.org/10.1145/3613904.3642910`.

[22] Lijun Li, Linrui Tian, Xindi Zhang, Qi Wang, Bang Zhang, Liefeng Bo, Mengyuan Liu, and Chen Chen. Renderih: A large-scale synthetic dataset for 3d interacting hand pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 20395–20405, 2023.

[23] Yilin Liu, Shijia Zhang, and Mahanth Gowda. Neuropose: 3d hand pose tracking using emg wearables. In *Proceedings of the Web Conference 2021*, pages 1471–1482, 2021.

[24] Yiyue Luo, Chao Liu, Young Joong Lee, Joseph DelPreto, Kui Wu, Michael Foshey, Daniela Rus, Tomás Palacios, Yunzhu Li, Antonio Torralba, et al. Adaptive tactile interaction transfer via digitally embroidered smart gloves. *Nature communications*, 15(1):868, 2024.

[25] Eric Marchand, Hideaki Uchiyama, and Fabien Spindler. Pose estimation for augmented reality: a hands-on survey. *IEEE transactions on visualization and computer graphics*, 22(12): 2633–2651, 2015.

[26] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In *Proceedings of the IEEE conference on computer vision and pattern Recognition*, pages 5079–5088, 2018.

[27] Yaru Niu, Yunzhe Zhang, Mingyang Yu, Changyi Lin, Chenhao Li, Yikai Wang, Yuxiang Yang, Wenhao Yu, Tingnan Zhang, Zhenzhen Li, et al. Human2locoman: Learning versatile quadrupedal manipulation with human pretraining. In *Robotics: Science and Systems*, 2025.

[28] Myungsun Park, Taejun Park, Soah Park, Sohee John Yoon, Sumin Helen Koo, and Yong-Lae Park. Stretchable glove for accurate and robust hand pose reconstruction based on comprehensive motion data. *Nature communications*, 15(1):5821, 2024.

[29] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9826–9836, 2024.

[30] Yuzhe Qin, Wei Yang, Binghao Huang, Karl Van Wyk, Hao Su, Xiaolong Wang, Yu-Wei Chao, and Dieter Fox. Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system. *arXiv preprint arXiv:2307.04577*, 2023.

[31] Santosh Rath. Hand kinematics: application in clinical practice. *Indian journal of plastic surgery*, 44(02):178–185, 2011.

[32] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. URL `https://arxiv.org/abs/2408.00714`.

[33] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), November 2017.

[34] Romer Rosales, Vassilis Athitsos, Leonid Sigal, and Stan Sclaroff. 3d hand pose reconstruction using specialized mappings. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 1, pages 378–385. IEEE, 2001.

[35] Christina Salchow-Hömmen, Leonie Callies, Daniel Laidig, Markus Valtin, Thomas Schauer, and Thomas Seel. A tangible solution for hand motion tracking in clinical applications. *Sensors*, 19(1):208, 2019.

[36] Sasha Salter, Richard Warren, Collin Schlager, Adrian Spurr, Shangchen Han, Rohin Bhasin, Yujun Cai, Peter Walkington, Anuoluwapo Bolarinwa, Robert J Wang, et al. emg2pose: A large and diverse benchmark for surface electromyographic hand pose estimation. *Advances in Neural Information Processing Systems*, 37:55703–55728, 2024.

[37] Kyungjin Seo, Junghoon Seo, Hanseok Jeong, Sangpil Kim, and Sang Ho Yoon. Posture-informed muscular force learning for robust hand pressure estimation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[38] Yuto Shibata, Yutaka Kawashima, Mariko Isogawa, Go Irie, Akisato Kimura, and Yoshimitsu Aoki. Listening human behavior: 3d human pose estimation with acoustic signals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13323–13332, 2023.

[39] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1145–1153, 2017.

[40] Jiaqi Sun, Guangda Liu, Yubing Sun, Kai Lin, Zijian Zhou, and Jing Cai. Application of surface electromyography in exercise fatigue: a review. *Frontiers in Systems Neuroscience*, 16:893275, 2022.

[41] Anilkumar Swamy, Vincent Leroy, Philippe Weinzaepfel, Fabien Baradel, Salma Galaaoui, Romain Brégier, Matthieu Armando, Jean-Sebastien Franco, and Grégory Rogez. Showme: Robust object-agnostic hand-object 3d reconstruction from rgb video. *Computer Vision and Image Understanding*, 247:104073, 2024.

[42] Arvin Tashakori, Zenan Jiang, Amir Servati, Saeid Soltanian, Harishkumar Narayana, Katherine Le, Caroline Nakayama, Chieh-ling Yang, Z Jane Wang, Janice J Eng, et al. Capturing complex hand movements and object interactions using machine learning-powered stretchable smart textile gloves. *Nature Machine Intelligence*, 6(1):106–118, 2024.

[43] Gyan Tatiya, Jonathan Francis, Ho-Hsiang Wu, Yonatan Bisk, and Jivko Sinapov. Mosaic: Learning unified multi-sensory object property representations for robot learning via interactive perception. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 15381–15387. IEEE, 2024.

[44] Stephen Tyree, Jonathan Tremblay, Thang To, Jia Cheng, Terry Mosier, Jeffrey Smith, and Stan Birchfield. 6-dof pose estimation of household objects for robotic manipulation: An accessible dataset and benchmark. In *International Conference on Intelligent Robots and Systems (IROS)*, 2022.

[45] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *International Journal of Computer Vision*, 118:172–193, 2016.

[46] Ryan J Visee, Jirapat Likitlersuang, and Jose Zariffa. An effective and efficient method for detecting hands in egocentric videos for rehabilitation applications. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(3):748–755, 2020.

[47] Rong Wang, Wei Mao, and Hongdong Li. Deepsimho: Stable pose estimation for hand-object interaction via physics simulation. *Advances in Neural Information Processing Systems*, 36: 79685–79697, 2023.

[48] Rong Wang, Wei Mao, and Hongdong Li. Interacting hand-object pose estimation via dense mutual attention. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 5735–5745, 2023.

[49] Shiyang Wang, Xingchen Wang, Wenjun Jiang, Chenglin Miao, Qiming Cao, Haoyu Wang, Ke Sun, Hongfei Xue, and Lu Su. Towards smartphone-based 3d hand pose reconstruction using acoustic signals. *ACM Transactions on Sensor Networks*, 20(5):1–32, 2024.

[50] Shiyang Wang, Henglin Pu, Qiming Cao, Wenjun Jiang, Xingchen Wang, Tianci Liu, Zhengxin Jiang, Hongfei Xue, and Lu Su. Ram-hand: Robust acoustic multi-hand pose reconstruction using a microphone array. In *Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems*, pages 130–143, 2025.

[51] Yilin Wen, Hao Pan, Lei Yang, Jia Pan, Taku Komura, and Wenping Wang. Hierarchical temporal transformer for 3d hand pose estimation and action recognition from egocentric rgb videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21243–21253, 2023.

[52] Min-Yu Wu, Pai-Wen Ting, Ya-Hui Tang, En-Te Chou, and Li-Chen Fu. Hand pose estimation in object-interaction based on deep learning for virtual reality applications. *Journal of Visual Communication and Image Representation*, 70:102802, 2020.

[53] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10974, 2019.

[54] Yufei Ye, Poorvi Hebbar, Abhinav Gupta, and Shubham Tulsiani. Diffusion-guided reconstruction of everyday hand-object interaction clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 19717–19728, 2023.

[55] Zhao-Heng Yin, Changhao Wang, Luis Pineda, Francois Hogan, Krishna Bodduluri, Akash Sharma, Patrick Lancaster, Ishita Prasad, Mrinal Kalakrishnan, Jitendra Malik, et al. Dexteritygen: Foundation controller for unprecedented dexterity. *arXiv preprint arXiv:2502.04307*, 2025.

[56] Masairo Yoshikawa, Masahiko Mikawa, and Kazuyo Tanaka. Hand pose estimation using emg signals. In *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4830–4833. IEEE, 2007.

[57] Tianhong Catherine Yu, Guilin Hu, Ruidong Zhang, Hyunchul Lim, Saif Mahmud, Chi-Jung Lee, Ke Li, Devansh Agarwal, Shuyang Nie, Jinseok Oh, et al. Ring-a-pose: A ring for continuous hand pose tracking. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(4):1–30, 2024.

[58] Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. Deep learning-based human pose estimation: A survey. *ACM Computing Surveys*, 56(1):1–37, 2023.

[59] Lucas Alexandre Zick, Dieisson Martinelli, André Schneider de Oliveira, and Vivian Cremer Kalempa. Teleoperation system for multiple robots with intuitive hand recognition interface. *Scientific Reports*, 14(1):1–11, 2024.

[60] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *Proceedings of the IEEE international conference on computer vision*, pages 4903–4911, 2017.

[61] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 813–822, 2019.