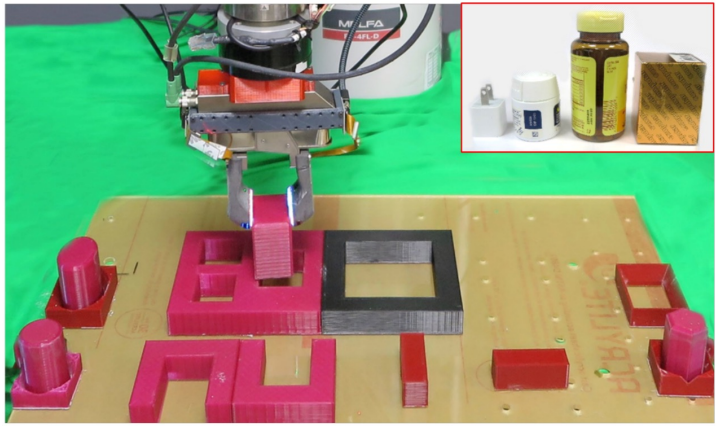
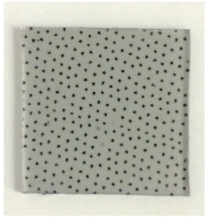
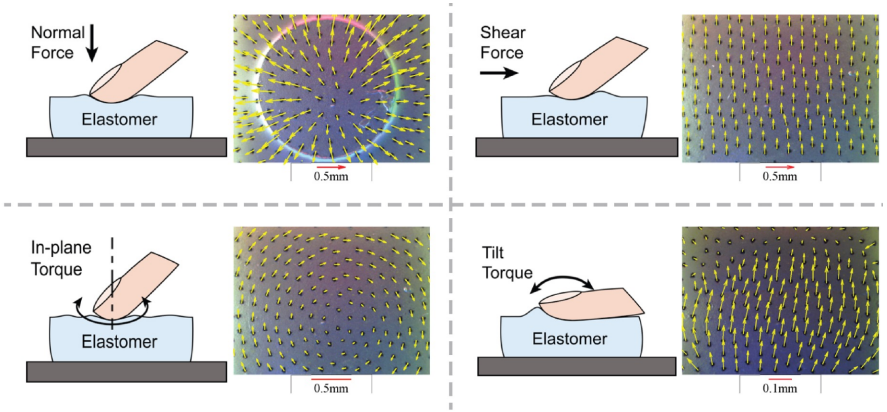
Tactile Sensing and Multimodal Perception

The Contact-Occlusion Problem





(a)



(b)

Same grasp. Fundamentally different information.

Key: this is a topological limitation - no camera resolution resolves contact occlusion.

Part 1

Why Touch?

The information geometry of contact

What Vision Cannot Tell You

Human fingertip: ~2,500 mechanoreceptive endings/cm²

SA-I (Merkel): sustained pressure, fine spatial detail

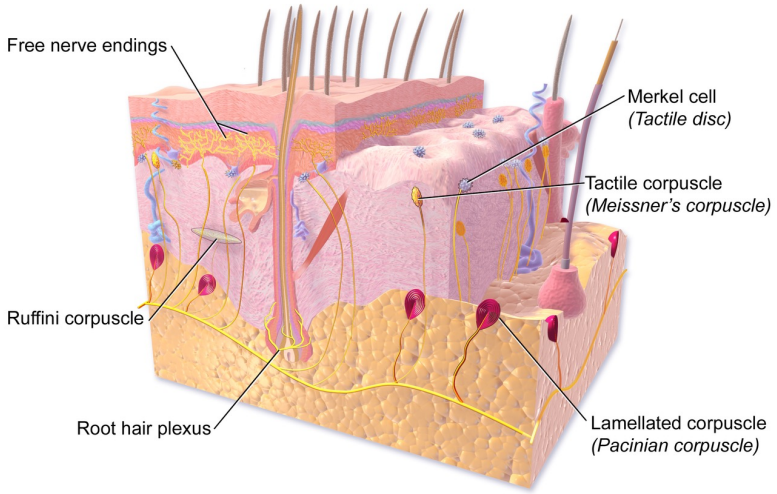
SA-II (Ruffini): skin stretch, finger pose

FA-I (Meissner): low-frequency vibration, slip onset

FA-II (Pacinian): high-frequency vibration >250 Hz

What vision cannot provide at contact:

1. Contact geometry at sub-millimeter resolution
2. Normal force distribution across the contact patch
3. Tangential force and incipient slip
4. Subsurface texture and material (hardness, compliance)



Tactile Receptors in the Skin

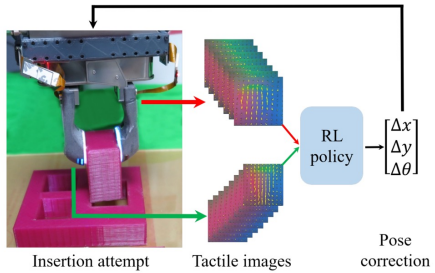
Key Insight

Vision and touch have complementary information geometries. Vision is high-bandwidth, non-contact, and global — it tells you where objects are in the scene. Touch is contact-local, force-sensitive, and material-discriminating — it tells you the mechanical state at the interface. No camera resolution can substitute for the second category, because the contact interface is physically occluded by the bodies in contact.

Three Tasks Where Touch is Decisive

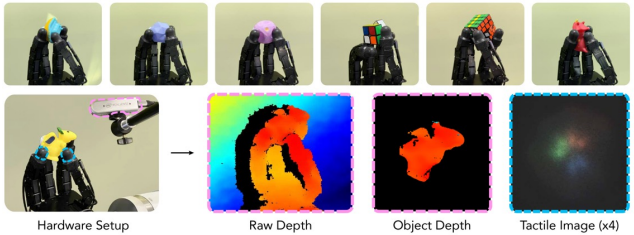
Insertion Tasks

Tolerances often <0.5 mm—below wrist-camera localization accuracy. Tactile signal: contact patch shape and lateral force gradient as peg approaches hole boundary. (Dong et al., ICRA 2021)



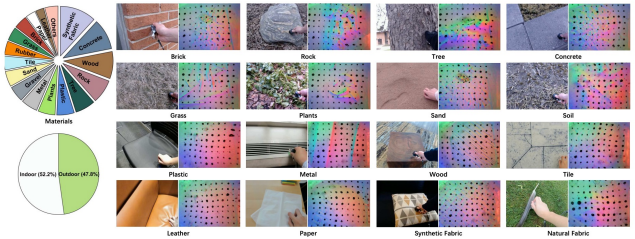
In-Hand Manipulation

Vision cannot see fingertip contact patches during in-hand rotation. Tactile sensing monitors contact patch migration toward the fingertip edge (pre-slip). (Yin et al. RSS 2023; Qi et al. CoRL 2023)



Texture and Material Discrimination

Visually identical objects can differ in surface texture, compliance, or hardness. Tactile sensing discriminates reliably, including for deformable objects (cable routing: She et al., IJRR 2021).



Part 2

Tactile Sensor Technologies

From resistive arrays to vision-based fingertips

Four Technology Families

Four transduction mechanisms convert mechanical contact to measurable signals. Optical (vision-based) sensors dominate current robot learning research.

Technology	Spatial Res.	Bandwidth	Contact Geometry	Cost	Fragility
Resistive	1–5 mm	~100 Hz	Pressure array	Low	Low
Capacitive	1–3 mm	~500 Hz	Pressure array	Medium	Medium
Piezoelectric	~1 mm	>1 kHz	Dynamic only (AC)	Medium	High
Optical (vision)	**<0.1 mm**	**~30–60 Hz**	**Full 2D geometry**	**Medium**	**Medium**

Note: Higher spatial resolution is not always better. Bandwidth and force sensitivity often matter more, depending on the task. Sensor selection should be task-driven.

GelSight Operating Principle

Symbol Definitions

Symbol	Domain	Meaning
$I_k(x,y)$	$\mathbb{R}\geq 0$	Pixel intensity at (x,y), light k
$\rho(x,y)$	$\mathbb{R}\geq 0$	Surface albedo at (x,y)
l_k	$\mathbb{R}^3$, $\|l_k\|=1$	Light direction unit vector
$n(x,y)$	$\mathbb{R}^3$, $\|n\|=1$	Surface normal at (x,y)
K	$\mathbb{Z}\geq 1$	No. of illumination directions

EQ1 Lambertian reflectance

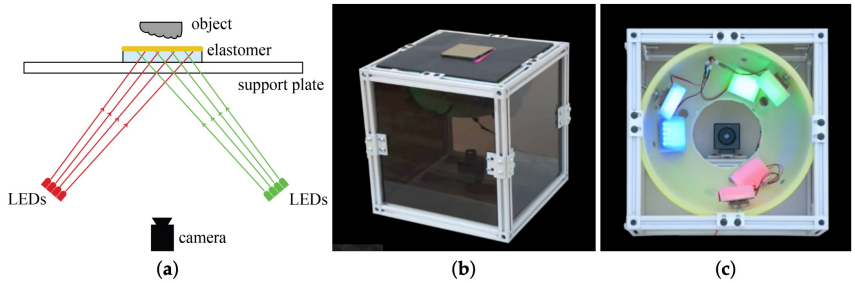
$$I_k(x,y) = \rho(x,y) \cdot \big(\mathbf{l}_k \cdot \mathbf{n}(x,y)\big)$$

EQ2 Stacked system (K directions)

$$\mathbf{I}(x,y) = \rho(x,y)\, L\, \mathbf{n}(x,y), \quad \mathbf{I} \in \mathbb{R}^N,\ L \in \mathbb{R}^{N\times 3}$$

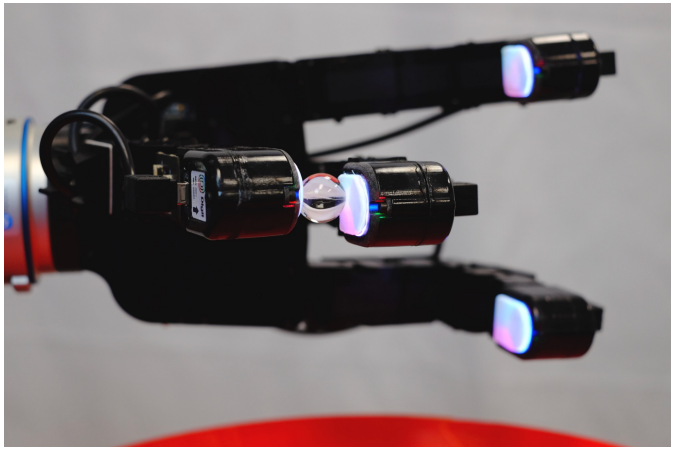
EQ3 Least-squares normal recovery

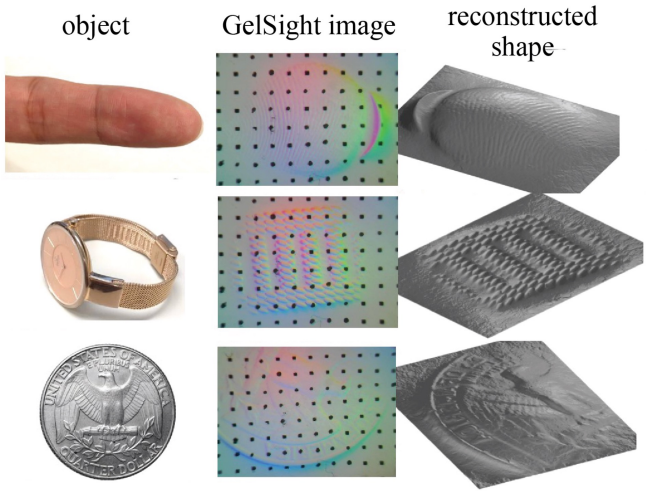
$$\hat{\mathbf{n}}(x,y) = \frac{(L^\top L)^{-1} L^\top \mathbf{I}(x,y)}{\left\|(L^\top L)^{-1} L^\top \mathbf{I}(x,y)\right\|}$$



The GelSight Family Tree

Year	Sensor / Authors	Key contribution
2009	Johnson & Adelson (CVPR, MIT CSAIL)	First tabletop photometric stereo on elastomer. Not yet a fingertip.
2014	Li et al. (IROS, MIT)	First robotic fingertip for parallel-jaw grippers.
2017	**Yuan, Dong & Adelson (Sensors, MIT)**	**Standardized design + algorithm. Canonical course reference.**
2020	**Lambeta et al. — DIGIT (RA-L, Meta FAIR)**	**Compact (~25 mm); drives NeuralFeels, Sparsh, 3D-ViTac.**
2022	Taylor et al. (ICRA, MIT + CMU)	GelSlim 3.0 — flat planar form factor, strip LEDs.
2025	Bhirangi et al. (ICRA, Cornell)	AnySkin — replaceable skins, magnetic attachment.

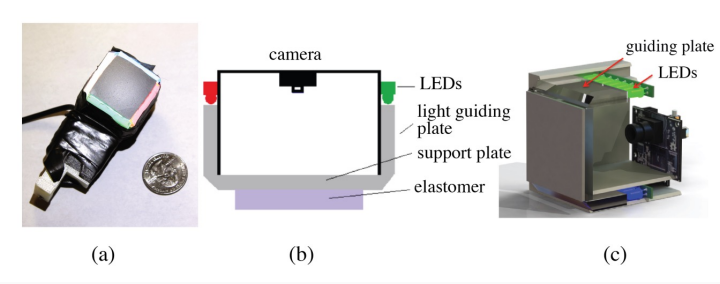




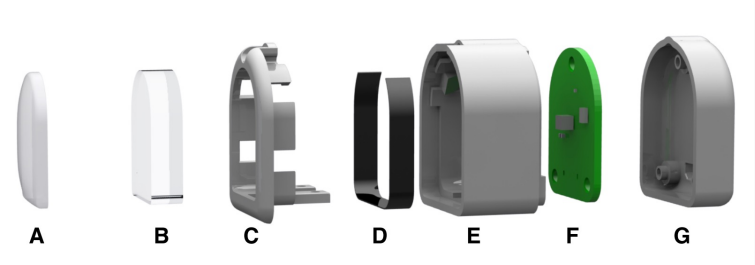
GelSight Hardware Evolution

Four generations of the GelSight design paradigm — each sharing the same photometric stereo pipeline despite very different physical form factors.

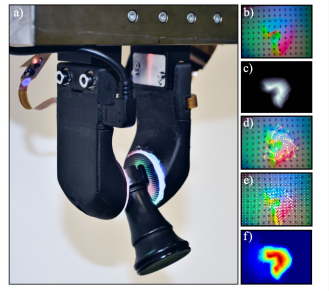
Yuan, Dong & Adelson, Sensors 2017

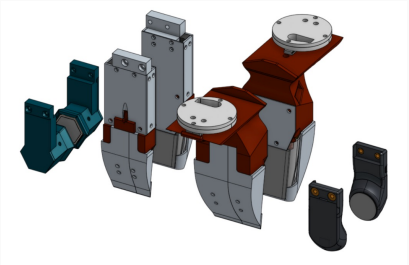


Lambeta et al., RA-L 2020 (DIGIT)

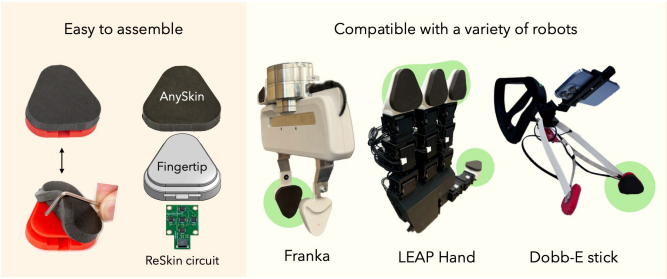


Taylor et al., ICRA 2022 (GelSlim 3.0)





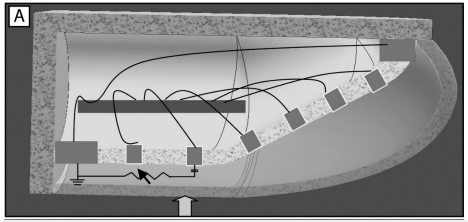
Bhirangi et al., ICRA 2025 (AnySkin)



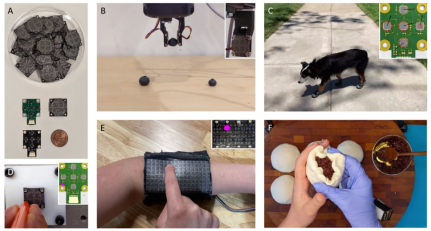
Beyond GelSight

Three non-GelSight sensors appear frequently in the literature — we need to recognize and distinguish them.

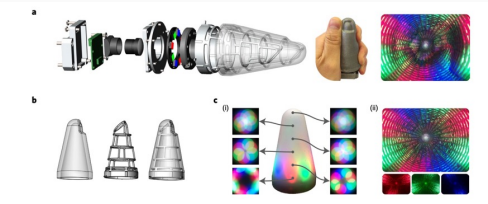
BioTac (SynTouch, 2008)



ReSkin (Bhirangi et al., CoRL 2021)



Insight (Sun et al., Nat. MI 2022)



Principle	Fluid-filled silicone	Magnetic silicone skin	Spherical; 360° camera
Output	22D vector (19 elec.)	3D flux per sensing node	2D image → 6-axis F/T
Best for	Compliance / impedance	~$6/sensor; no recal.	Non-planar 3D contact

Learning Connection

Every tactile sensor in this slide produces a different data modality — a 22-dimensional vector (BioTac), a 3D magnetic flux map (ReSkin), or a high-resolution 2D image (GelSight family). Identify the sensor first when reading any tactile manipulation paper. The sensor determines what information is available and whether learned representations transfer across sensors.

Sensor Technology Summary

Pipeline bridge: from physical contact to the contact geometry consumed in later architectures. Each sensor family enters the pipeline at a different transduction stage.

Sensor family	Transduction	Raw signal	Processing	Output to Part 3
Resistive / Capacitive	ΔR or ΔC under pressure	Scalar array (N×1)	Interpolation, calibration	Pressure map
Piezoelectric	Charge ΔQ under strain	AC voltage array	High-pass filter + envelope	Dynamic signal only
Magnetic (ReSkin)	Field distortion ΔB	3D flux vector/node	Calibrated lookup table	3D contact force / pos.
Optical / Vision (GelSight family)	**Elastomer surface normal change**	**Color camera image**	**Photometric stereo (EQ1–3)**	**Normal map + height map h(x,y)**

→ Part 3 (next section) processes the GelSight-family output: normal map → height map → contact mask.

Part 3

Classical Tactile Signal Processing

From raw sensor output to contact geometry

Photometric Stereo and Height Map Integration

Symbol	Domain	Meaning
h(x,y)	$\mathbb{R}$	Surface height at pixel (x,y)
p(x,y)	$\mathbb{R}$	x-gradient: p = −n_x/n_z
q(x,y)	$\mathbb{R}$	y-gradient: q = −n_y/n_z
Δ	—	Laplacian operator

EQ4a Height map integration

$$\min_h \iint \left[\left(\frac{\partial h}{\partial x} - p \right)^2 + \left(\frac{\partial h}{\partial y} - q \right)^2 \right] dx\, dy$$

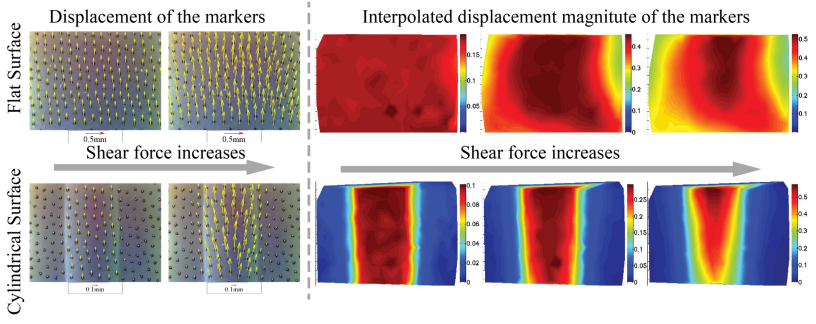
EQ4b Poisson form

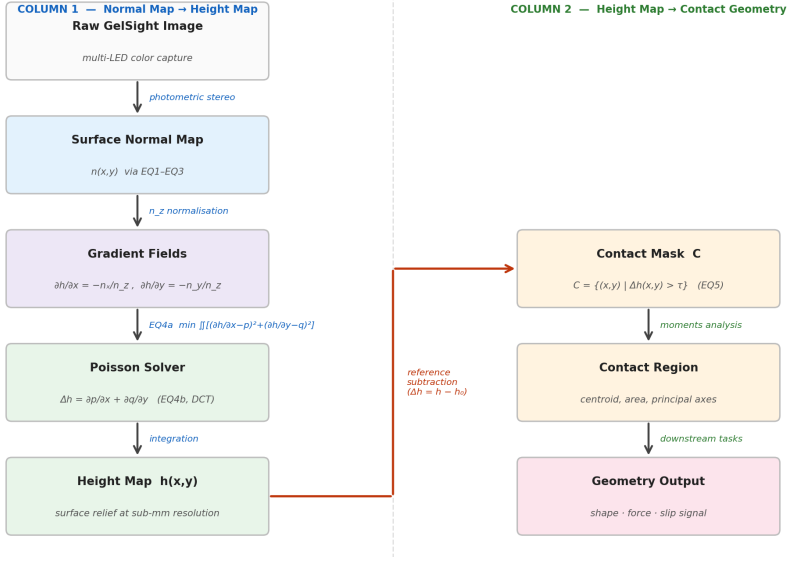
$$\Delta h = \frac{\partial p}{\partial x} + \frac{\partial q}{\partial y}$$

Symbol	Domain	Meaning
h_0(x,y)	$\mathbb{R}$	Reference height map
Δh(x,y)	$\mathbb{R}$	Δh = h(x,y) − h_0(x,y)
τ	$\mathbb{R}$>0	Contact threshold (~0.05 mm)
C	—	Contact mask (where Δh>τ)

EQ5 Contact mask

$$\mathcal{C} = \{(x, y) \mid \Delta h(x, y) > \tau\}$$





Force Estimation and Slip Detection

Symbol	Domain	Meaning
$f \in \mathbb{R}^3$	Force	Normal + tangential force vector
$\varphi(\Delta h)$	$\mathbb{R}^d$	Feature vector from Δh
W, b	—	Regression weights + bias

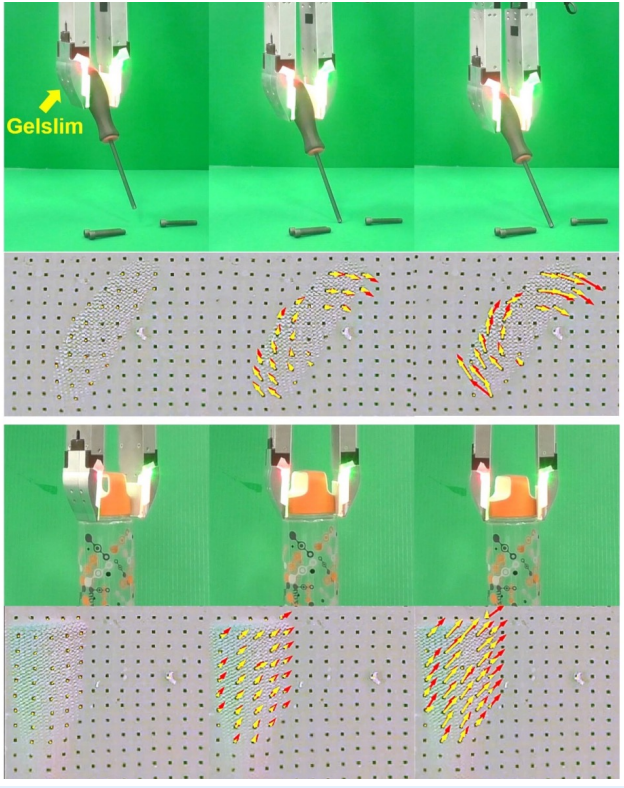
EQ6 Linear force regression

$$\hat{\mathbf{f}} = W\,\psi(\Delta h) + \mathbf{b}$$

Symbol	Domain	Meaning
I(x,y,t)	$\mathbb{R}$	Tactile image at time t
∇I	$\mathbb{R}^2$	Spatial image gradient
S(t)	$\mathbb{R}\geq 0$	Slip signal at time t
$\|\cdot\|_s$	—	Frobenius norm (spatial)

EQ7 Slip signal — frame-to-frame gradient change

$$S(t) = \left\| \nabla I(\cdot,\cdot,t) - \nabla I(\cdot,\cdot,t-1) \right\|_{\mathrm{F}}$$



Learning Connection

The classical pipeline — photometric stereo, Poisson integration, contact masking, force regression, slip detection — produces exactly the labeled signals that self-supervised tactile representation learning uses as pre-training targets.

Part 4

Tactile Representation Learning

From task-specific networks to general-purpose encoders

The Representation Learning Problem

A raw GelSight image (640×480) has ~300k values; the contact patch covers only 5–15% of pixels. What should a network extract from this input?

Task-Specific Supervised (2018–2021)

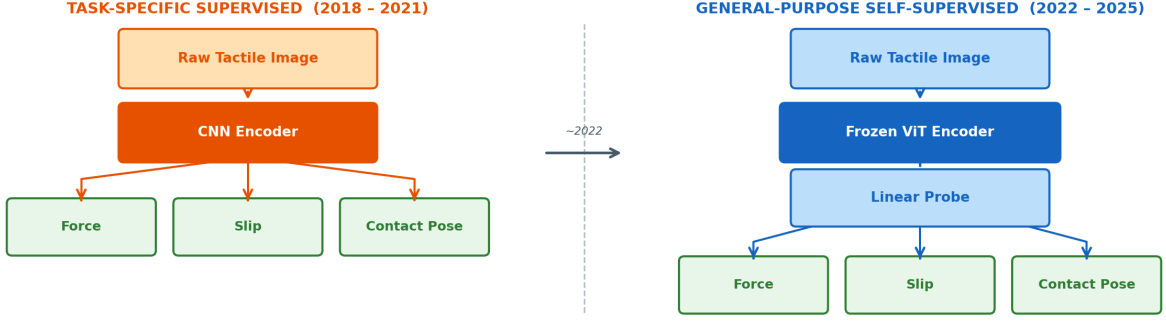
Train a CNN on (image → label) pairs for one output: force, slip, or contact pose.

Generalizes poorly. New labeled dataset required for every new task and sensor.

General-Purpose Self-Supervised (2022–2025)

Pre-train a ViT on large unlabeled tactile data (masked / contrastive / predictive objectives).

Freeze encoder; attach a lightweight head per task. One encoder, many tasks, no labels.



Key Insight

The evolution from task-specific tactile networks to general-purpose pre-trained encoders mirrors the trajectory of visual representation learning: from AlexNet trained per-task to frozen ViTs adapted via linear probing.

Self-Supervised Pre-Training for Touch

T-Dex — Guzey et al., CoRL 2023

Contrastive pre-training on robotic play data: random exploration with a DIGIT-equipped fingertip.

Positive pairs: temporally nearby frames. Negatives: distant frames.

Key result: enables dexterous manipulation from far fewer demonstrations than raw image conditioning. Backbone: ViT-S.

Sparsh — Higuera et al., CoRL 2024 (Meta FAIR)

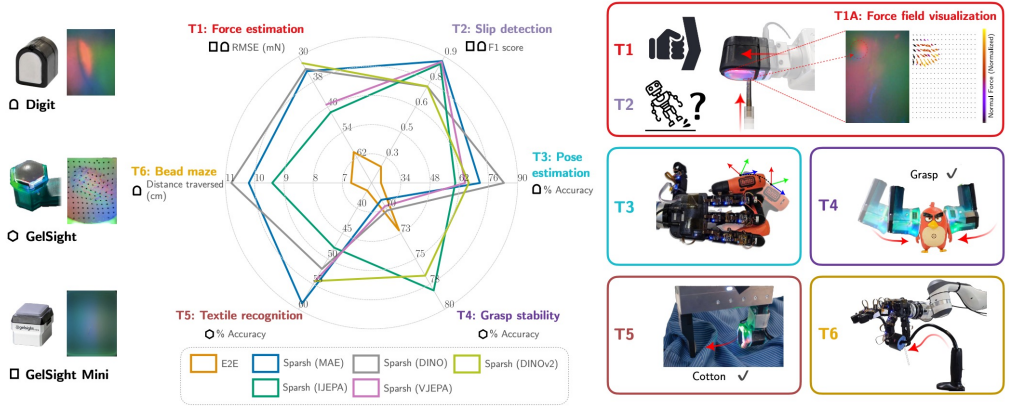
Pre-trains MAE, DINO, and IJEPA encoders on 460k+ tactile images from DIGIT and GelSight.

Introduces TacBench: benchmark across contact localization, force estimation, slip detection, and pose estimation.

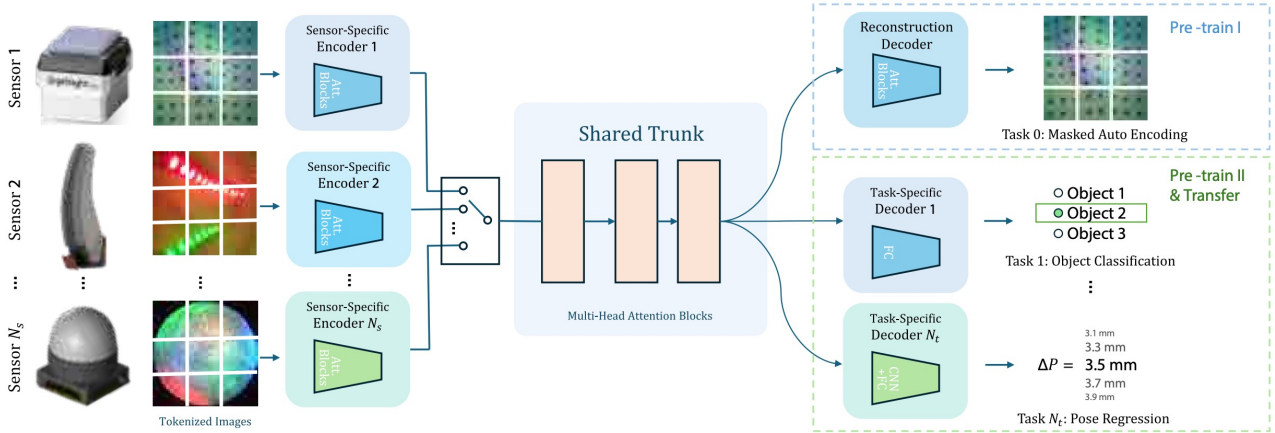
Key result: IJEPA outperforms MAE and DINO. Encoder publicly released.

Architecture note:

Both methods use a standard ViT backbone. Tactile images are 2D arrays with local patch structure — exactly the inductive bias ViT handles well. The self-supervised objectives are identical to those for RGB images; only the domain changes.



Cross-Sensor Generalization — T3



T3 (Transferable Tactile Transformers)

Zhao et al., CoRL 2024 — MIT CSAIL

Problem: A GelSight encoder does not transfer to DIGIT images because illumination geometry, gel material, and camera perspective differ. Sensor-specificity is a major practical obstacle.

Architecture: Sensor-specific lightweight heads (one per sensor type) feed into a shared ViT trunk. Heads handle domain alignment; trunk learns sensor-agnostic contact representations.

FoTa dataset: Foundation Tactile — 3M+ datapoints from 13 sensors across 11 manipulation tasks. An order of magnitude larger than any prior tactile dataset.

Key result: T3 trunk features transfer zero-shot to unseen sensors, requiring only fine-tuning of the sensor-specific head. Exceeds task-specific baselines trained on full target-domain data.

Cross-Modal Alignment — UniTouch

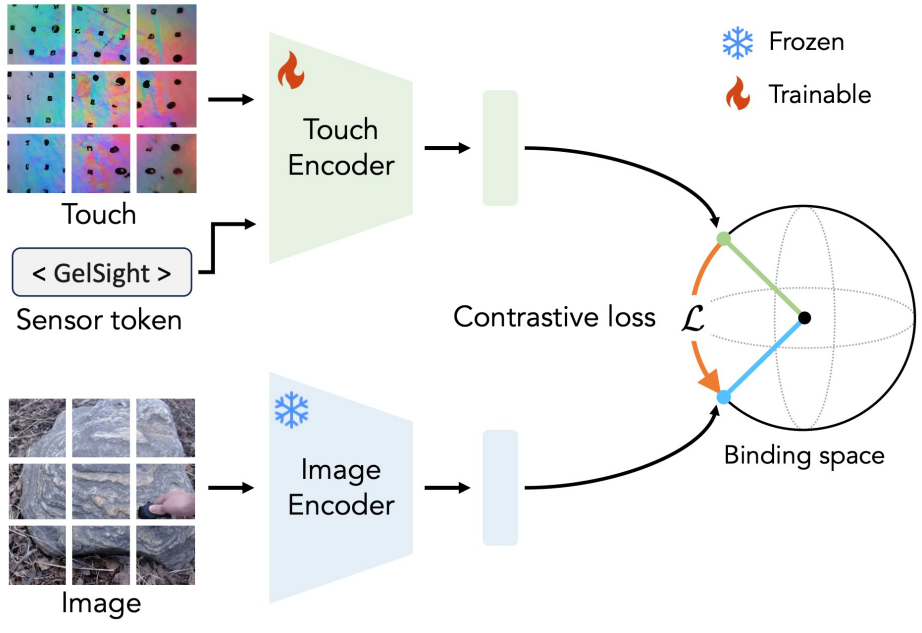
UniTouch — Yang et al., CVPR 2024

University of Michigan

Idea: CLIP aligned images and text via 400M pairs. UniTouch applies the same strategy to pair tactile observations with visual observations of the same surface.

Training: Paired (tactile image, RGB image of same surface) from the Touch and Go dataset. The tactile encoder is trained to align its output to the frozen CLIP image embedding.

Results: Zero-shot material classification via CLIP text embeddings. Cross-modal retrieval: given a tactile observation, find visually similar surfaces.



Learning Connection

UniTouch demonstrates that the CLIP alignment strategy transfers from vision-language to touch-vision. This has a direct implication for Vision-Language-Action models: if tactile embeddings live in the same space as CLIP image embeddings, then VLAs pre-trained on vision and language could in principle consume tactile observations.

Part 5

Vision-Tactile Fusion for Robot Policies

Designing observation spaces that see and feel

The Fusion Design Space

Early Fusion	Mid-Level / Attended Fusion	Late Fusion / Modality-Conditioned Residual
Concatenate raw or lightly processed tactile and visual feature vectors before the policy. Simple to implement. Fails in practice: the two modalities have very different spatial scales. Naïve concatenation dilutes the tactile signal in high-dim. visual features.	Each modality processed by a modality-specific encoder into a fixed-dimension embedding. Fused via cross-attention or learned gating. Allows the network to learn which modality to attend to at each moment. The dominant approach in current tactile policies.	Separate policy paths for vision and touch. The tactile signal modulates the visual policy's action via a residual correction. Natural for extending a pre-trained visual policy with tactile feedback. Preserves visual policy performance while adding contact sensitivity.

Three questions that determine fusion architecture choice:

1. **Temporal relationship** Touch updates at contact events; vision updates continuously. The fusion layer must handle asynchronous inputs.
2. **Spatial relationship** Touch is contact-local and fingertip-fixed; vision is scene-global. A shared 3D representation (see 3D-ViTac) can bridge this gap.
3. **Pre-training availability** Frozen tactile encoder + frozen visual encoder + small fusion head is the most data-efficient regime when labeled data is scarce.

NeuralFeels — Visuotactile Object Tracking

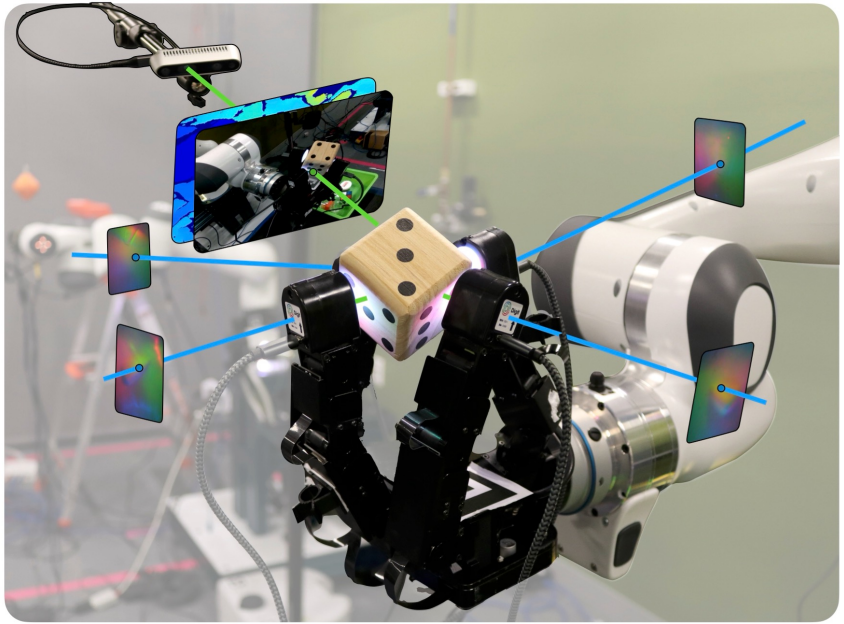
NeuralFeels — Suresh et al., Science Robotics 2024

CMU + Meta FAIR

Problem: Vision-only methods accumulate unbounded drift during long in-hand manipulation sequences because feature correspondences fail as the object moves.

Approach: Online neural field (NeRF-like implicit shape) maintained for the tracked object. RGB-D observations provide scene-level constraints via differentiable rendering. DIGIT tactile readings (photometric stereo → 3D contact patch) provide contact-local surface constraints — the recovered contact geometry must be consistent with the neural field at the contact location. Both modalities contribute gradients.

Key result: Tracks object pose over long manipulation sequences where vision-only methods drift. Touch provides absolute surface constraints that anchor the estimate.



3D-ViTac — Unified 3D Visuo-Tactile Policy

3D-ViTac — Huang et al., CoRL 2024

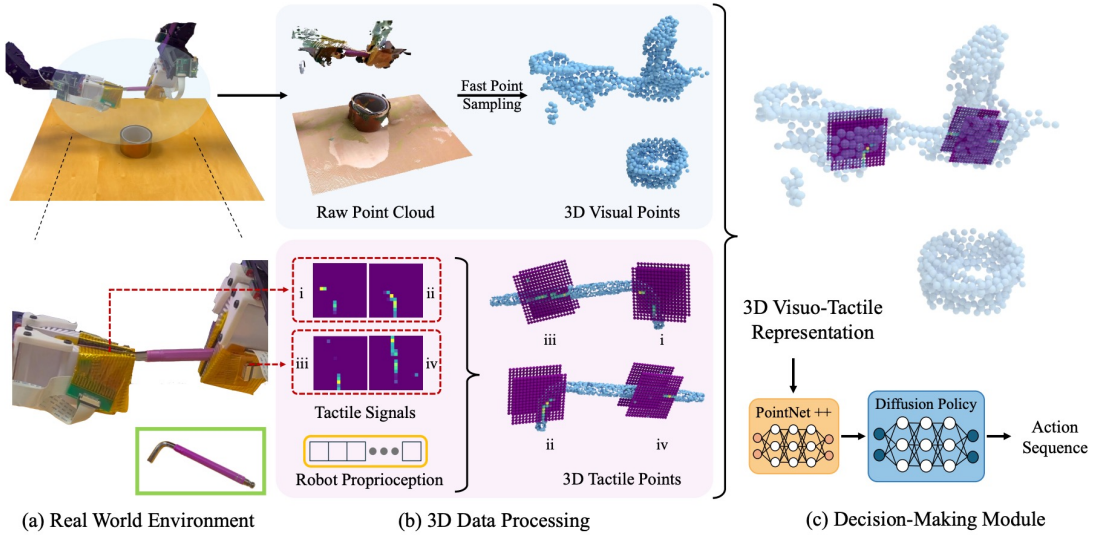
Step 1 — Visual: RGB-D from wrist + overhead cameras → 3D visual point cloud P_{vis} via depth unprojection.

Step 2 — Tactile: DIGIT readings → photometric stereo → 3D contact point cloud P_{ta}^{c} in fingertip frame → world frame via FK.

Step 3 — Unify: $P = P_{vis} \cup P_{ta}^{c}$ — a single unified 3D point cloud.

Step 4 — Policy: Diffusion Policy consumes the unified point cloud.

Key result: Substantially higher success rates than vision-only across 10 bimanual fine-grained tasks. The ablation shows the gain comes from 3D unification, not from adding tactile features in 2D image space.



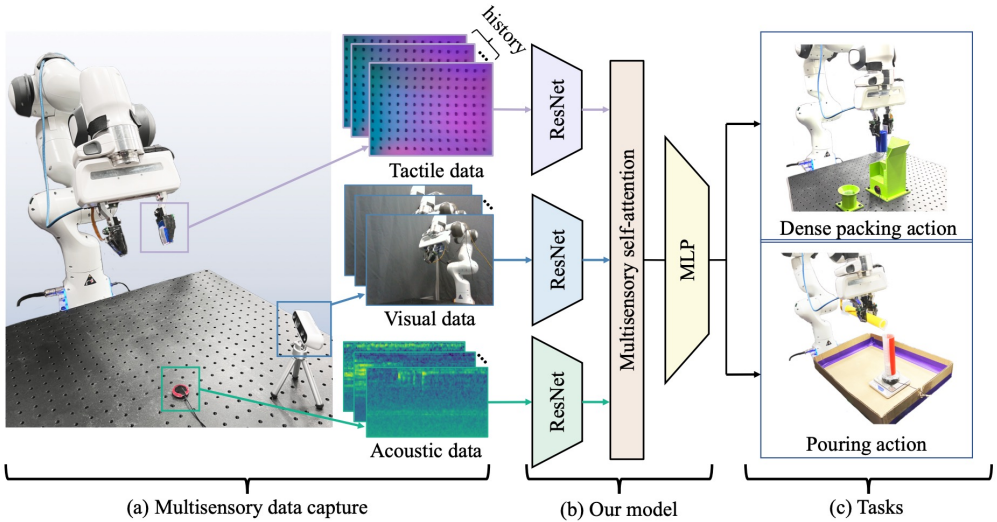
See, Hear, and Feel — Multi-Modal Attention

Li et al., CoRL 2022 — Stanford + MIT

Approach: Separate pre-trained encoders for each modality — ResNet (vision), VGG (audio spectrogram), CNN (tactile image) — produce fixed-dimension embeddings. These are treated as tokens and processed by a small Transformer.

Key result: On three tasks (cup stacking, spice bottle manipulation, peg insertion), the policy outperforms any single-modality baseline and any fixed-weight fusion. Attention weights are interpretable: during contact events, attention to the tactile token increases sharply.

Note: Modality weighting is an emergent learned behavior, not a hand-designed rule. This principle reappears at larger scale in Vision-Language-Action models, where attention across many input modalities is learned jointly.



In-Hand Manipulation — What Touch Enables

Touch Dexterity (Yin et al., RSS 2023) and RotateIt (Qi et al., CoRL 2023): general in-hand object rotation with three-fingered hand and DIGIT sensors per fingertip.

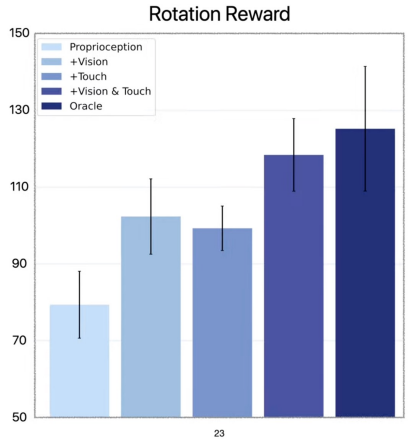
Task condition	Vision -only	Touch- only	Vision + Touch
Symmetrical objects	22%	61%	87%
Novel geometries	38%	49%	82%
After partial occlusion	15%	68%	79%

Key Insight

In-hand manipulation reveals an asymmetry in the roles of vision and touch: vision provides the global task context — where the object is, what orientation it needs to reach — while touch provides the local contact control signal — is the grasp stable, is slip imminent.

Key design point

The tactile signal serves as a **contact-state monitor**, not a controller. The policy learns when contact state suggests impending slip and preemptively adjusts fingertip forces. This requires temporal integration of tactile readings across the trajectory — a single-frame tactile observation is insufficient.



Part 6

Proprioception

The robot’s internal body sense

Proprioceptive Hardware

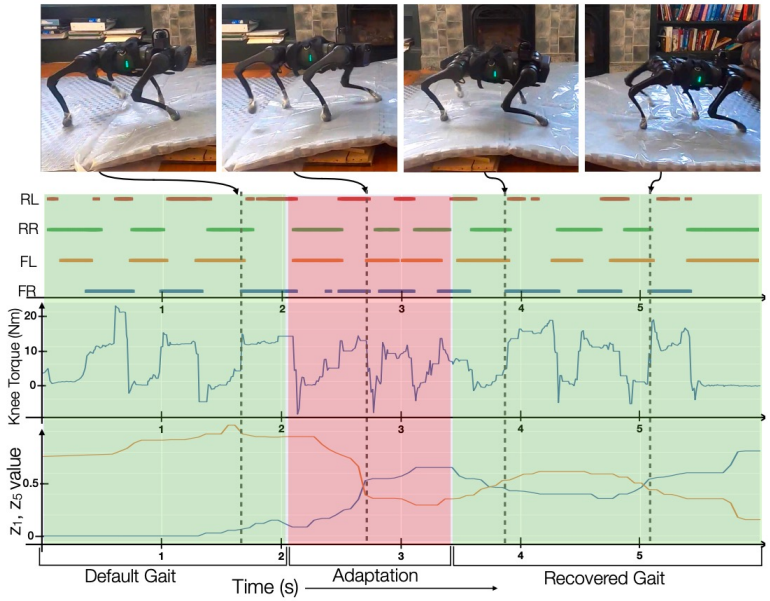
Always available — proprioception works in the dark, without visual targets, and without contact.

Joint encoders: Measure joint angle. Resolution 12–20 bit; latency ~1 ms. Velocity estimated by differentiation.

Joint torque sensors: Strain-gauge bridges measuring transmitted torque. Essential for impedance control and contact detection. Resolution ~0.01 Nm.

IMU: 3-axis accelerometer + gyroscope. 200–1000 Hz update rate. Critical for locomotion, drifts.

Component	Dimension	Update rate
Joint positions q	6	1000 Hz
Joint velocities q̇	6	1000 Hz
Joint torques τ	6	1000 Hz
End-effector pose (FK)	7 (pos + quat)	Derived
Wrist F/T	6	500 Hz
Total	**31**	**—**



Proprioception in Robot Learning Pipelines

Lee et al., Science Robotics 2020 — ANYmal Blind Locomotion

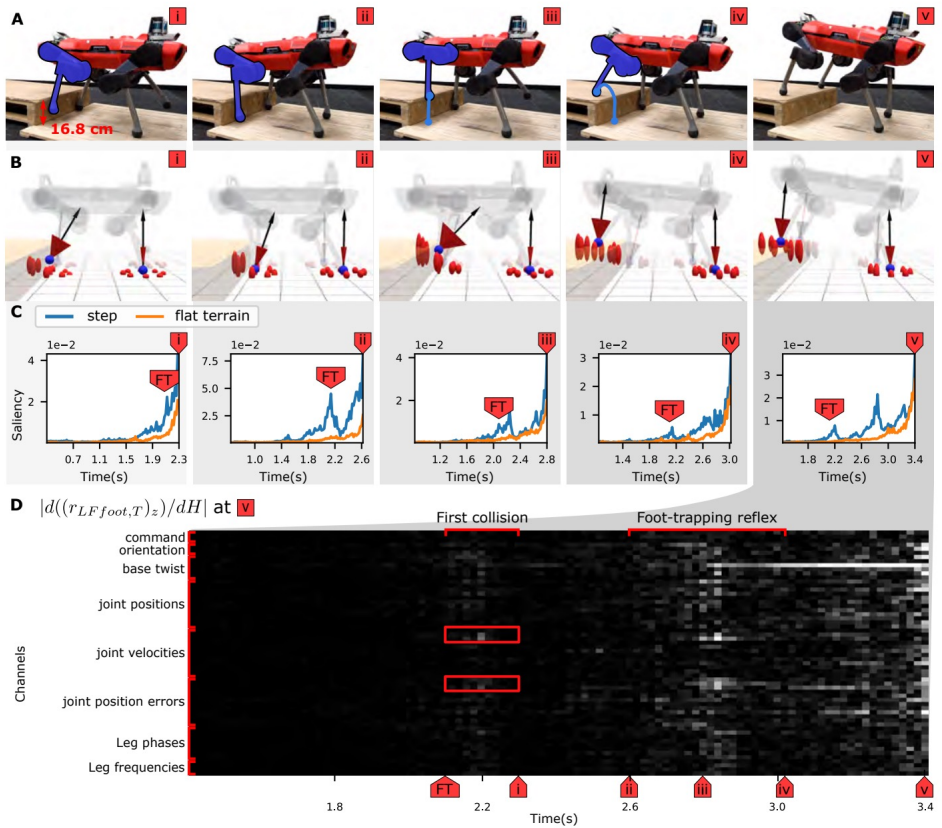
Trained purely on proprioceptive inputs (joint pos/vel/torque, IMU) — no cameras. Achieves robust locomotion over mud, snow, rubble, and stairs. Key lesson: proprioception alone carries sufficient state for high-performance locomotion.

RMA — Kumar et al., RSS 2021

Two-stage: base policy uses privileged environment state; adaptation module maps proprioceptive history (50 steps) → latent encoding matching Stage 1's privileged state. Key insight: proprioceptive history is an implicit sensor for ground friction, payload, slope.

HPT — Wang et al., NeurIPS 2024

Heterogeneous Pre-trained Transformers: modality-specific tokenizers feed a shared trunk. Key insight: heterogeneous tokenization outperforms naïve concatenation.



Part 7

Tactile Simulation and the Sim-to-Real Gap

Why simulating touch is harder than simulating vision

Tactile Simulators — Approaches and Trade-offs

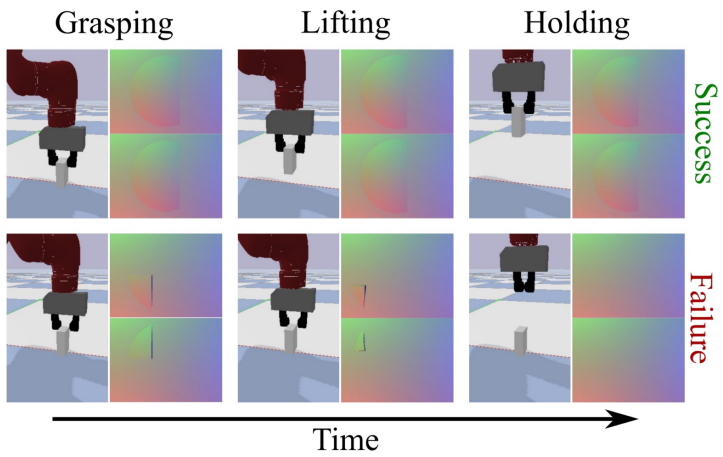
TACTO (Wang et al.): OpenGL rendering of GelSight/DIGIT appearance for a given contact mesh. Fast (~100 fps), optionally differentiable. Limitation: elastomer deformation approximated geometrically.

Taxim (Si & Yuan): Example-based. Interpolates from a database of real GelSight images at known depths. More accurate than TACTO in-distribution but does not generalize to novel geometries.

DiffTactile (Si et al.): Fully differentiable. FEM elastomer model + differentiable renderer. Gradients flow from rendered image through contact mechanics to policy. Enables gradient-based policy optimization through touch.

TacIPC (Du et al.): IPC-based elastomer simulation guaranteeing intersection-free deformation. More physically robust than DiffTactile but slower (~1–5 fps).

Property	Rendering-based (TACTO / Taxim)	Physics-based (DiffTactile / TacIPC)
Speed	Fast (~100 fps)	Slow (~1–10 fps)
Accuracy	Approximate (geometric elastomer)	Physically consistent (FEM)
Differentiable	Optional	Yes — end-to-end



Why Tactile Sim-to-Real is Hard

Challenge 1 — Elastomer Nonlinearity

The gel is a hyperelastic, viscoelastic material. Simple linear elasticity models underestimate force response at large deformations and miss rate-dependence. Domain randomization over elastic moduli partially compensates but does not capture creep and relaxation behaviors.

Challenge 2 — Illumination Drift

The GelSight reflective coating oxidizes over time, changing reflectance. A simulation calibrated to a new sensor mismodels a worn one. This is an ongoing per-sensor calibration problem, not a one-time fix.

Challenge 3 — Contact Stiffness Dynamics

High-frequency contact dynamics (micro-impacts, stick-slip transitions) occur at time scales most simulators cannot resolve at practical rates. Mitigation: learned latent projection (Narang et al., ICRA 2021) maps simulated signals into a latent space matching real signal distribution without closing the physical gap.

The Full Sensory Stack

Three qualitatively distinct modalities, each occupying a different niche in the information space of physical manipulation. They are not interchangeable.

Modality	What it measures	Available when	Key limitation
Vision	Scene geometry, object identity, pose	If lit and in frame	Occluded at contact
Touch	Contact geometry, force, slip, material	Only at contact	Sim-to-real gap
Proprioception	Body state, joint torques	Always	No environmental info

Policies that fuse all three modalities are more capable and more robust than policies relying on any one alone — but fusion architecture must respect the modality differences rather than treating all inputs as equivalent feature vectors.