



Sensor-Invariant Tactile Representation

Harsh Gupta*, Yuchen Mo*, Shengmiao Jin, Wenzhen Yuan



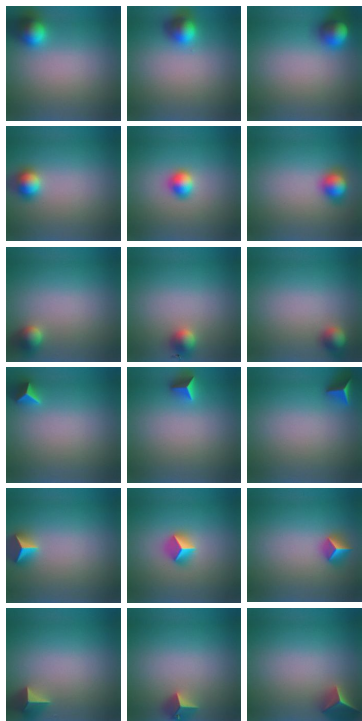
Tactile sensors are different

- Lack of standardized formats and protocols.
- Significant differences in physical structure and illumination methods across sensors.
- Transferring learning-based methods require labor-intensive data collection for each sensor.



Tactile sensors are not reliable

- Even two GelSight Mini sensors have sensor domain differences.
- The exact same sensor may also shift over time. Minor wear, such as scratches on the gel surface, degrade model performance.

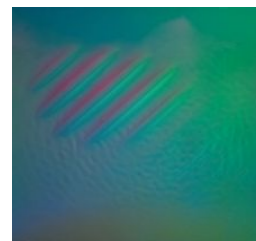
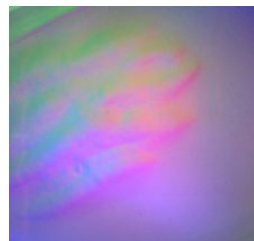
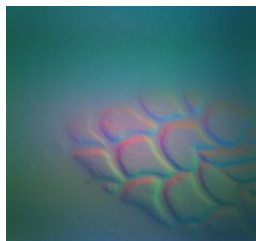


Why is calibration important?

- For GelSight-like sensors, the RGB values at each pixel correspond to the local surface gradient, enabling the reconstruction of the entire contact surface through integration

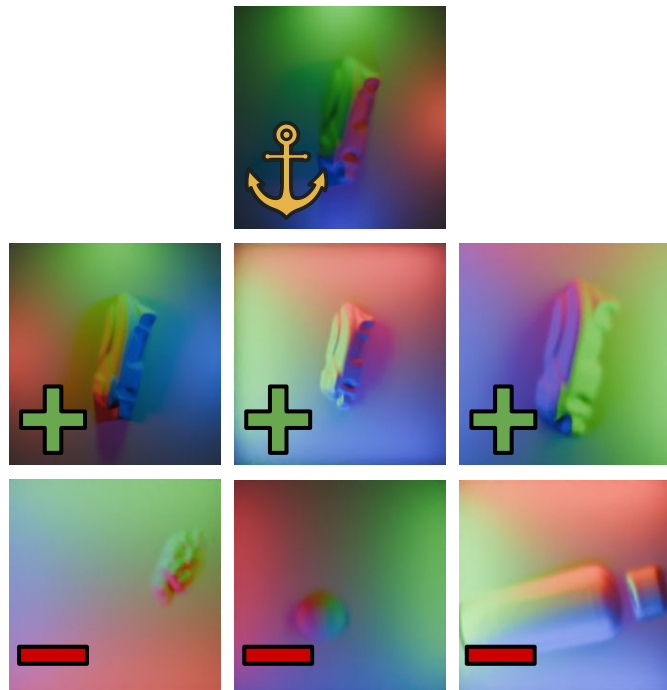
But this approach relies on perfect sensors

- Sensors come in all shapes and sizes and exhibit artifacts and biases in different ways



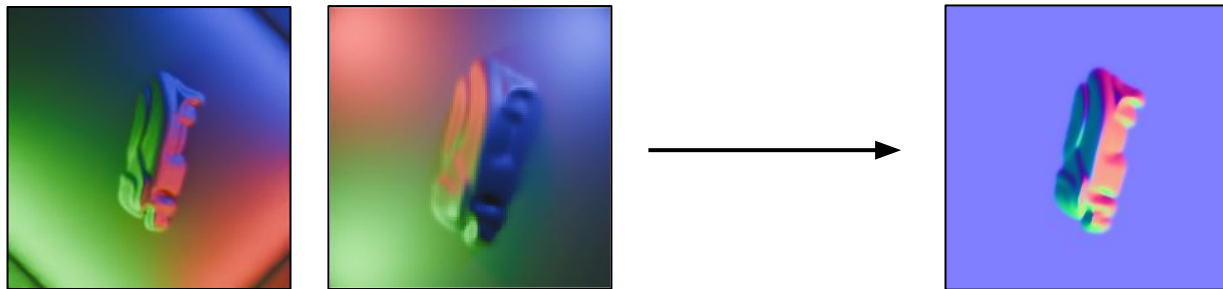
Preserve global features

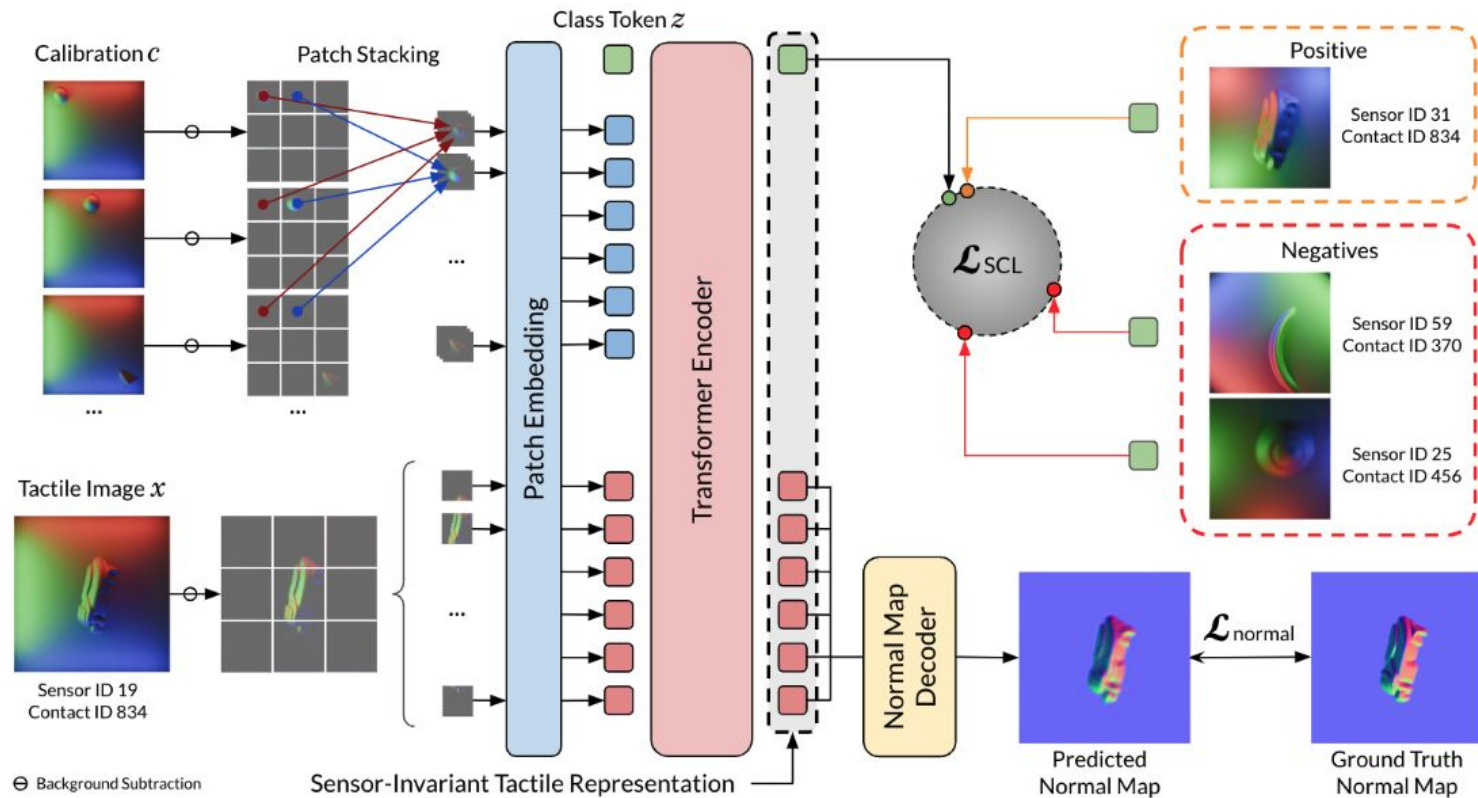
- We use contrastive learning in order to preserve important global features that are required for downstream tasks.
- Using our simulated dataset, we create positive and negative pairs for each sample in order to push similar samples closer in latent space, while pushing dissimilar samples farther apart.
- Our pairs are set up in order to align the representations to be invariant to the sensor



Preserve contact rich information

- Normal maps provide important cues about contact geometry and textures that are important in many downstream tasks.
- Normal maps are more dependent on the contact than the sensor. So they serve as a great supervising signal.
- We use a normal map decoder to ensure that these features are preserved in our latent space for downstream tasks.



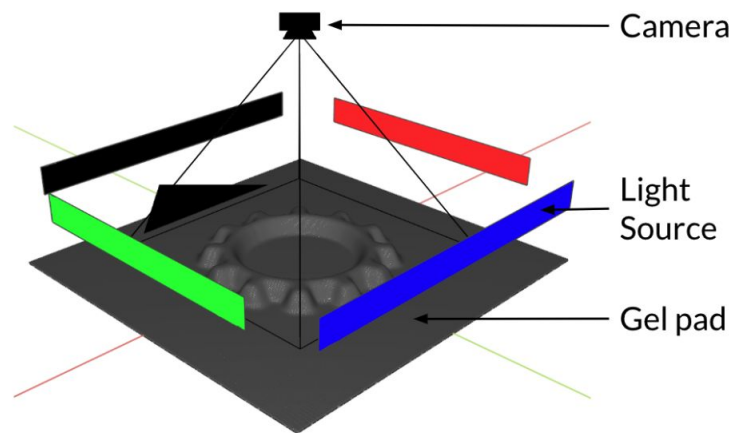




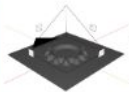
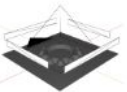


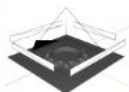



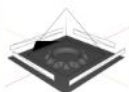



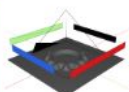
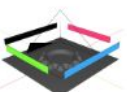
Training dataset











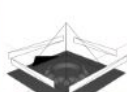



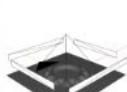

- **Simulated dataset:** We generate and deploy 100 random sensor configurations, each with its own set of calibration images. ~1M samples

We randomize a number of parameters

1. Color variance
2. Light type (point/area)
3. Light angle
4. Gel stiffness
5. Gel specularity
6. Camera FOV and scale

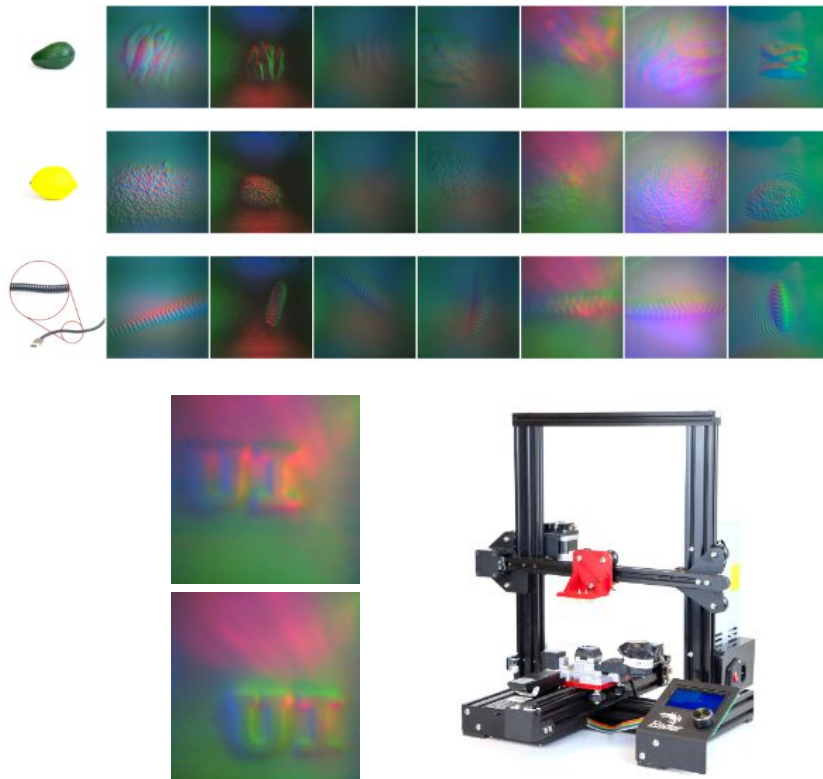


Parameter	Lower bound	Upper bound	Lower bound vis.	Upper bound vis.	Lower bound env.	Upper bound env.
Light shape	point	area				
Light orientation	sides	corners				
Light angle	5°	30°				
Light color	rand	rand				

Parameter	Lower bound	Upper bound	Lower bound vis.	Upper bound vis.	Lower bound env.	Upper bound env.
Gel stiffness	low	high				
Gel specularity	low	high				
Camera FOV	40°	90°				
Sensing area	4cm ²	16cm ²				

Evaluation datasets

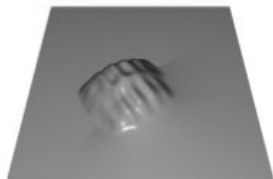
- **Classification dataset:** We collect a 16 object dataset across 7 sensors. These 7 sensors consist of 2 sets. Intra-sensor set of 4 GS Mini with varied gel pads. Inter-sensor set of 4 separate sensors: GS Mini, GS Wedge, GS Hex, DIGIT. ~Total 112K samples
- **Pose estimation:** We record the location of probes as the indent a sensor. Using 2 images we calculate the difference between them. We collect this over 6 probes across the 4 inter-sensor set. ~Total 24K samples.



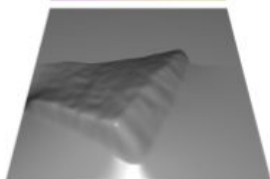
Qualitative results



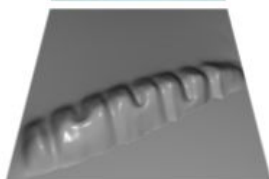
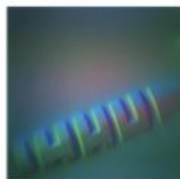
Simulation 1



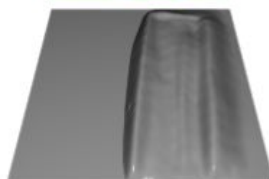
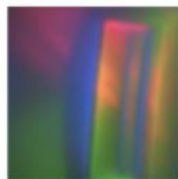
Simulation 2



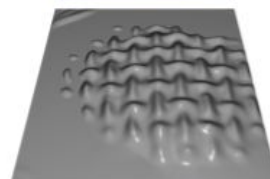
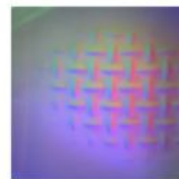
GelSight Mini



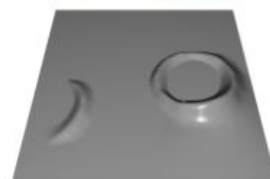
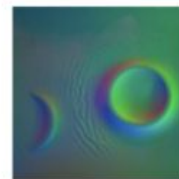
DIGIT

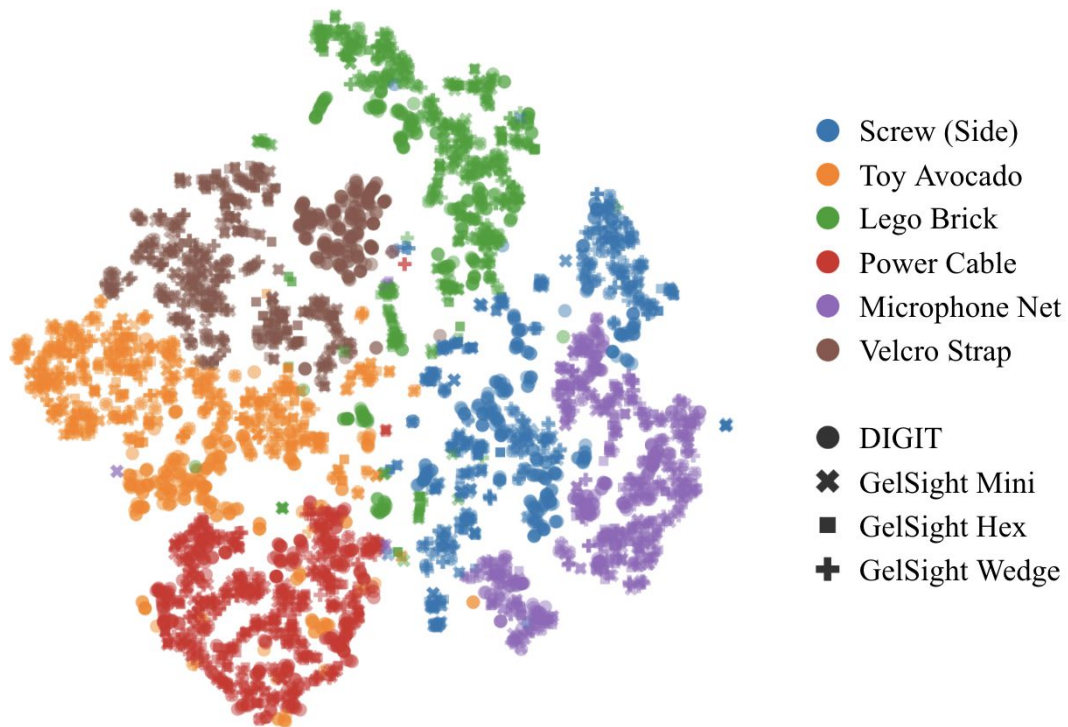


GelSight Hex



GelSight Wedge





Classification task

Method	Intra-sensor set \uparrow	Inter-sensor set \uparrow	Wedge-Mini \uparrow	No transfer \uparrow
ViT-Base Scratch	36.90 \pm 22.19	24.02 \pm 14.83	52.56 \pm 4.95	96.76 \pm 1.41
ViT-Base Pre-trained	73.22 \pm 22.42	48.10 \pm 22.82	76.28 \pm 17.06	99.01 \pm 1.14
ViT-Large Pre-trained	78.38 \pm 17.79	54.34 \pm 23.04	79.04 \pm 16.44	99.44 \pm 0.43
T3-Medium	38.66 \pm 20.63	— —	17.02 \pm 8.55	93.77 \pm 2.87
UniT	46.39 \pm 23.30	— —	— —	92.53 \pm 4.19
SITR (Ours)	90.23 \pm 8.16	81.94 \pm 12.92	90.80 \pm 2.85	99.72 \pm 0.22

Table 1: Results of object classification accuracy on 16 classes for model transfer and no-transfer performance. We report the mean and standard deviation of transfer accuracy percent among the sensor sets specified. Random guess classification accuracy corresponds to 6.67%.

Pose estimation task

Method	Inter-sensor set ↓	Wedge-Mini ↓	No transfer ↓
ViT-Base Scratch	1.63 ± 0.20	1.69 ± 0.13	0.56 ± 0.02
ViT-Base Pre-trained	1.58 ± 0.22	1.65 ± 0.13	0.49 ± 0.01
ViT-Large Pre-trained	1.49 ± 0.25	1.45 ± 0.01	0.50 ± 0.02
T3-Medium	— —	1.7 ± 0.07	0.51 ± 0.02
SITR (Ours)	0.80 ± 0.21	0.62 ± 0.11	0.51 ± 0.01

Table 2: Results of pose estimation with 6 objects. We report the mean and standard deviation of transfer pose estimation root mean square error (RMSE) in *mm* among the sensor sets specified. Random guess pose estimation RMSE corresponds to $2.52mm$.