Supplementary Materials for Visual-Tactile Sensing for In-Hand Object Reconstruction

Wenqiang Xu[*1,2], Zhenjun Yu[*1], Han Xue[1], Ruolin Ye[3], Siqiong Yao[1], Cewu Lu[§1,2]
[1]Shanghai Jiao Tong University [2]Shanghai Qi Zhi institute [3]Cornell University
[1]{vinjohn, jeffson-yu, xiaoxiaoxh, yaosiqiong, lucewu}@sjtu.edu.cn
[3]ry273@cornell.edu

1. The relationship between tactile pixel, depth and object local geometry

As we mentioned in our paper, we predict the depth from tactile images using a 2D-UNet, and then project to point clouds in the space. Examples are shown in Fig. 1. The local point clouds represent exactly the object geometry which contact with the sensor.

The reason why a local tactile pixel and the geometry have such a strong relationship is because the tactile image represents the deformation of the gel.

The gel deformation can be modeled by:

$$\boldsymbol{F} = \boldsymbol{K}\boldsymbol{U},$$

where $\boldsymbol{U} = (\delta_x^1, \delta_y^1, \delta_z^1, ..., \delta_x^n, \delta_y^n, \delta_z^n)$ is the displacement of n gel mesh vertices. $\boldsymbol{F} = (f_x^1, f_y^1, f_z^1, \ldots, f_x^n, f_y^n, f_z^n)$ are the external forces applied to the gel vertices. $\boldsymbol{K}$ is a $3n \times 3n$ matrix which represents the stiffness of the gel. K is determined by 2 parameters, namely Young's Modulus and Poisson's ratio. These parameters can be obtained by calibrating the real-world gel materials.

In the simulation, we do not actually deform the gel according to the contact force but simply insert the contact geometry. It is the same practice in TACTO [1]. It means we let $\boldsymbol{K} = \boldsymbol{I}$. This simplification will **not** influence the geometry estimation, as it only matters to the force estimation. And in real world, if we can estimate the geometry, we can recover the force by the calibrated $\boldsymbol{K}$.

Consider the camera distance to the original gel surface is $\boldsymbol{D}$, the depth for local geometry is $\boldsymbol{D}_{depth} = \boldsymbol{D} - \boldsymbol{U}$. And we can interprete the tactile image as the direct observation of $\boldsymbol{U}$. Thus, each pixel in tactile image has a corresponding physics meaning of local displacement $\boldsymbol{U}$, geometry $\boldsymbol{D}_{depth}$ and contact force $\boldsymbol{F}$.

2. Fingertip poses estimation by OpenCV

As shown in Fig. 2, we attach ArUco markers on the two finger tips with DIGIT, and predict the finger tip poses by a simple OpenCV algorithm. We follow the calibration process described in the official tutorial of OpenCV. Since it is a standard pipeline, we ask the reader to the official tutorial[1].

3. Examples on Real Dataset (Lock)

We have selected 3 different object from **Lock** category in the AKB-48 benchmark to test our model. We also adopt an Intel realsense L515 camera mounted on a fixed tripod to capture the depth observation, and we attach two DIGIT sensors to the hand and detect the sensor poses according to markers attached on tips. We augment the tactile images in simulation to match real RGB distribution by adding noise and adjusting contrast, and crop the input point cloud to remove noises.

We demonstrate examples in Fig. 2. Although the results are not as good as in the simulation dataset, the overall shape of locks has been reconstructed, and the effect of tactile signals can still be seen.

4. Supplementary Qualitative Results

In this section, we demonstrate more visualization results for VTacO and VTacOH. We present more examples for procedural tactile embedding in Fig. 3 and the comparison with visual-only and Visual & Tactile method in Fig. 4.

We gradually conduct more grasps on rigid objects to better reconstruct the textures and structures. We can see the structure of the small scissor has been completed during procedural grasping. As we mentioned in our main paper, the deformed regions of deformable objects such as **Bottles** and **Boxes** have been successfully reconstructed during multiple contacts. Such property can also be used for reconstructing plastic object with sustainable deformation.

We compare the visual-only method and VTacO in Fig. 4. Objects from the three categories in AKB-48 benchmarks

[1]https://docs.opencv.org/4.x/d5/dae/tutorial_aruco_detection.html

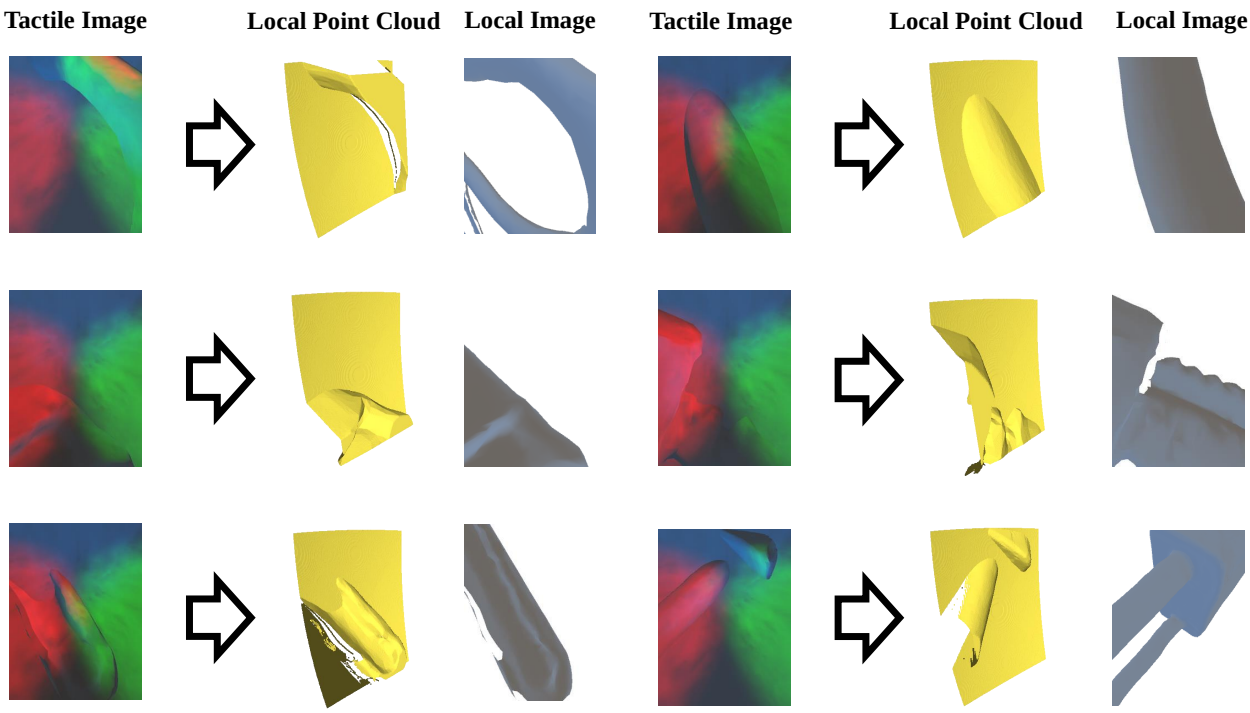

Figure 1. Local point cloud from depth prediction

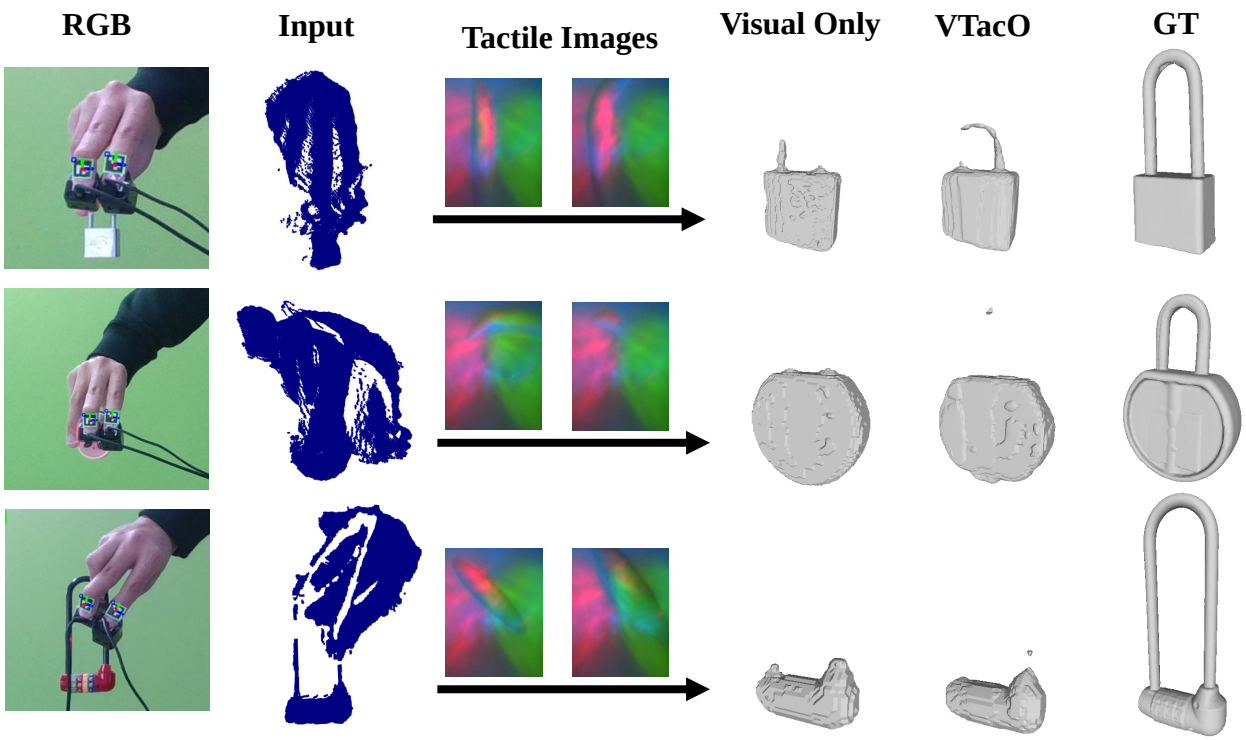

Figure 2. Examples of real dataset with **lock**

are better reconstructed with our method VTacO. As we illustrated, the introduction of tactile images helps to complete the whole structure, and somtimes it can remove the part of over-modeling, such as the second folding rack.

5. Broader Impact

- Our method will not directly or indirectly facilitate injury to living beings, because we only use real life objects for reconstruction.
- We will not cause serious accidents, open security vulnerabilities, nor will we collect user data or deploy surveillance when conducting experiments in real-world environments.
- We respect human rights in all aspects in our experiments.
- The experiements we conduct have no detrimental effect on people's livelihood or economic security.
- Since we use objects in real life to conduct real world experiments, we have no harm to the environment.
- To our knowledge, our method can not be used to deceive people in ways that cause harm.

References

[1] Shaoxiong Wang, Mike Lambeta, Po-Wei Chou, and Roberto Calandra. Tacto: A fast, flexible, and open-source simulator for high-resolution vision-based tactile sensors. *IEEE Robotics and Automation Letters*, 7(2):3930–3937, 2022. 1

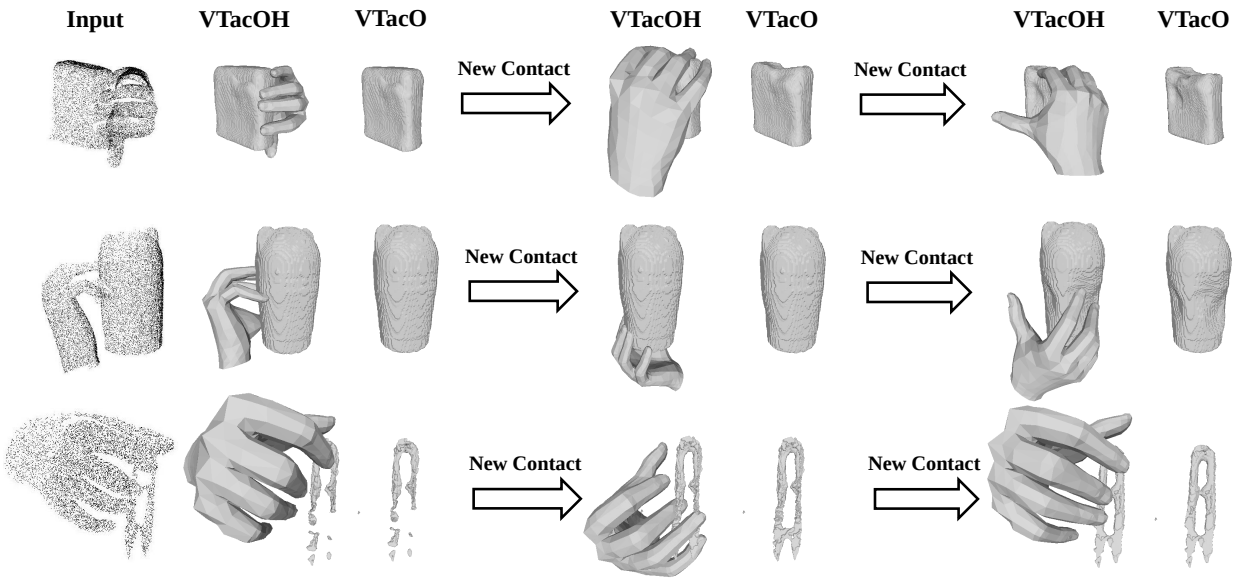

Figure 3. Procedural Tactile Embedding

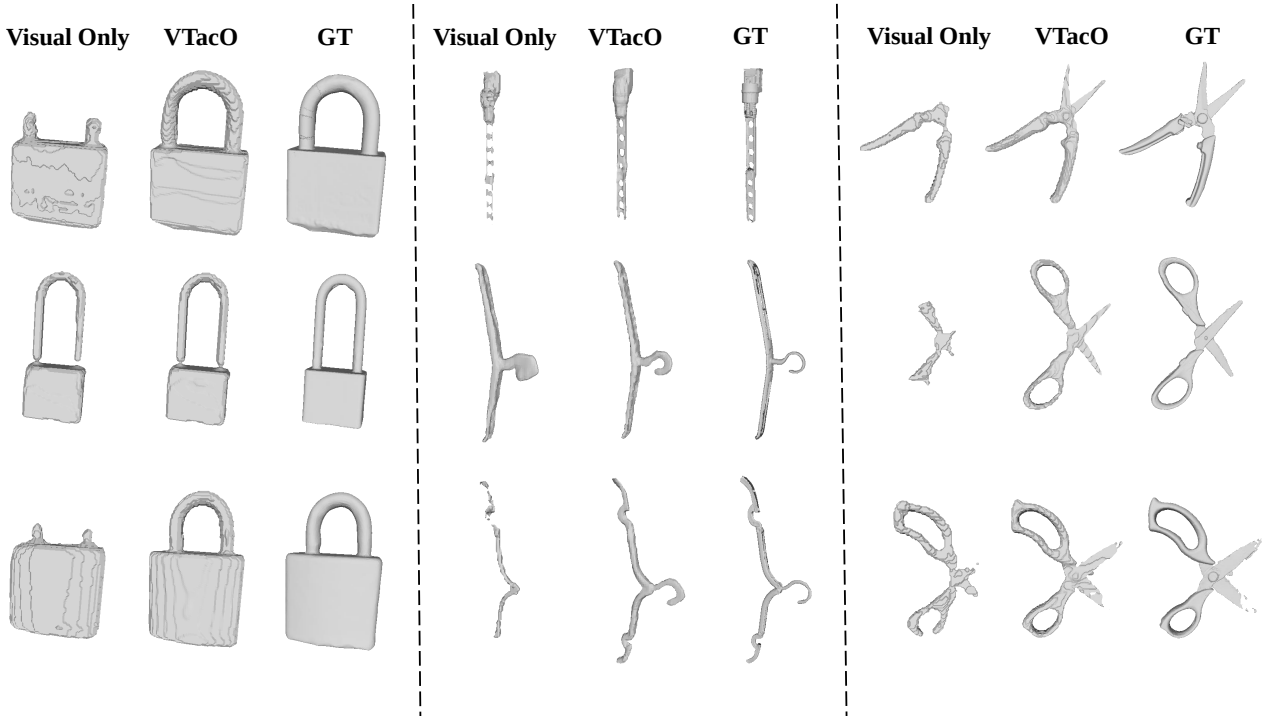

Figure 4. Comparison of results between pure visual method and VTacO