ShapeMap 3-D: Efficient shape mapping through dense touch and vision

Sudharshan Suresh*[1], Zilin Si*[1],
Joshua G. Mangelson[2], Wenzhen Yuan[1], and Michael Kaess[1]

Abstract— Knowledge of 3-D object shape is of great importance to robot manipulation tasks, but may not be readily available in unstructured environments. While vision is often occluded during robot-object interaction, high-resolution tactile sensors can give a dense local perspective of the object. However, tactile sensors have limited sensing area and the shape representation must faithfully approximate non-contact areas. In addition, a key challenge is efficiently incorporating these dense tactile measurements into a 3-D mapping framework. In this work, we propose an incremental shape mapping method using a GelSight tactile sensor and a depth camera. Local shape is recovered from tactile images via a learned model trained in simulation. Through efficient inference on a spatial factor graph informed by a Gaussian process, we build an implicit surface representation of the object. We demonstrate visuo-tactile mapping in both simulated and real-world experiments, to incrementally build 3-D reconstructions of household objects.

I. INTRODUCTION

For general-purpose manipulation in unstructured scenes, robots must have accurate understanding of object properties. In particular, knowledge of 3-D shape and its uncertainty enables a breadth of downstream tasks like grasping, dexterous manipulation, and non-prehensile actions. Agents in household or warehouse environments may encounter apriori unknown objects, which they must reconstruct on the fly.

Vision and depth-based 3-D perception is well-studied [1], but can often fail in the context of manipulation. During some interactions, we only partially observe the scene due to self-occlusion, occlusion from clutter, and fixed viewpoint. Also, visual sensing is degraded by poor illumination, limited range, and ambiguities from transparent or specular objects.

Studies show humans can optimally fuse touch and vision to reconstruct shape [2], reinforcing their complementarity. Vision gives coarse global context, while touch gives precise local information. The development of vision-based tactile sensing [3, 4, 5, 6, 7, 8, 9, 10], like the GelSight [4], has led to renewed interest in the shape mapping problem. Fusing both modalities requires globally integrating tactile signals at the distal end, joint kinematics, and vision.
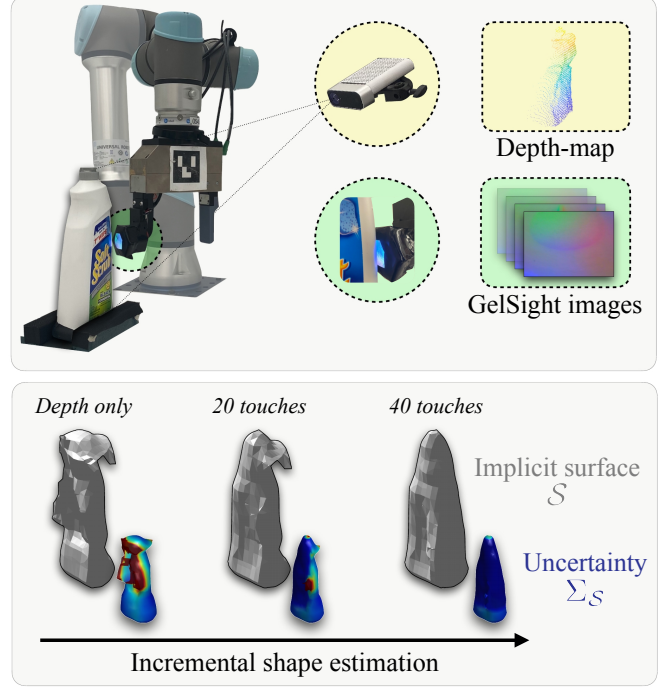
Fig. 1: We perform incremental 3-D shape mapping with a vision-based tactile sensor, GelSight, and an overlooking depth-camera. We combine multi-modal sensor measurements into our Gaussian process spatial graph (GP-SG), for efficient incremental mapping. The depth-camera gives us an occluded noisy estimate of 3-D shape, after which we sequentially add tactile measurements as Gaussian potentials into our GP-SG. The tactile measurements are recovered from GelSight images via a learned model trained in simulation. The results demonstrate accurate implicit surface reconstruction and uncertainty prediction for interactive perception tasks.

Vision-based touch has higher spatial acuity than point-contact or tactile arrays, which lends itself to 3-D reconstruction [11, 12, 13, 14]. A key challenge is to efficiently incorporating these dense measurements into a 3-D mapping framework. Moreover, the tactile sensor's coverage is limited by its size and durability, and cameras only provide partial visibility of the object. It's desired that a shape representation can faithfully approximate regions lacking measurements.

In this paper, we propose a framework that incrementally reconstructs tabletop 3-D objects from a sequence of tactile images and a noisy depth-map (Figure 1). We leverage optical tactile simulation to learn local shape from GelSight-object interactions. We represent 3-D shape as a signed distance function (SDF) sampled from a Gaussian process (GP), and re-formulate shape mapping as probabilistic inference on a spatial graph. We show that visuo-tactile measurements can be incorporated into a factor graph as local Gaussian potentials. This affords efficient access to the implicit sur-

face and SDF uncertainty. We present both simulated and real experiments, generating reconstructions of global shape with trajectories of limited sensor coverage. Specifically, our contributions include:

(1) Accurate recovery of local shape from touch learned via tactile simulation of GelSight-object interactions,
(2) Incremental shape mapping through efficient inference in our Gaussian process spatial graph (GP-SG),
(3) Evaluation of visuo-tactile shape mapping on our `YCBSight-Sim` and `YCBSight-Real` datasets.

II. RELATED WORK

A. Tactile sensing and local shape

For vision-based tactile sensors, photometric stereo [15] has been widely used to reconstruct local shape [4, 16, 17]. The approach maps image intensities to gradients via a lookup table, and integrates the gradients to obtain a height-map. However, this method does not consider spatial position in the calibration, and leads to large variance around the boundary of the sensor. Later works learn these gradients directly from tactile images, via either a multilayer perceptron [10] or pix2pix networks [18]. Alternatively, end-to-end learning [12, 19] from a limited set of real-world tactile interactions can directly provide heightmaps. Tactile simulation [20, 21, 22, 23] allows us to scale supervised-learning to a wide range of objects and ground-truth.

B. Visuo-tactile shape perception

Global information from vision has complemented low-resolution touch in a multi-modal setting [24, 25, 26, 27]. Wang et al. [11] use monocular shape completion augmented with GelSight readings. However they rely primarily on the visual shape prediction, and tactile sensing serves as a refinement step. Smith et al. [13, 14] demonstrate a learned perception model on simulated datasets, to predict local mesh deformations via high-resolution touch and filling-in through vision. The context of our work resembles those of [11] and [13], with partial vision and high-dimensional touch. Our contributions differ from these methods as we (i) perform incremental inference on the measurement stream, and (ii) do not rely on data-driven shape priors.

C. Gaussian processes and graphs

We wish to faithfully approximate non-contact regions, capture surface uncertainty, and probabilistically handle measurement noise. Gaussian process implicit surfaces (GPIS) [28] showcase these properties and have found preference in manipulation research—over point-clouds [12] and other parametric methods [29]. The GPIS considers the object's SDF magnitude and gradient as a GP, conditioned on noisy sensor measurements. This has been successfully applied to both passive [30, 31] and active 3-D reconstruction [24, 32, 33, 34] with low-resolution tactile data. We extend these ideas, scaling them to a stream of high-dimensional touch measurements for incremental shape reconstruction.

The key challenge, especially for GelSight point-clouds, is that GPs scale poorly due to matrix inversion costs. In the SLAM community, common approximations include local GPs [35, 36] and compact kernels [37]. These have further been incorporated into factor graphs [38] for trajectory estimation [39], target tracking [40], motion planning [41], elevation modeling [42], and planar mapping [43]. Inspired by these, our representation encodes GP potentials as local constraints in a spatial factor graph.

III. PROBLEM FORMULATION

We consider a robot arm with a GelSight tactile sensor interacting with an unknown 3-D object fixed on a tabletop. Given a sequence of images from the GelSight, robot kinematics, and depth-map from a depth-camera, we incrementally estimate the object's shape and signed distance function (SDF) uncertainty.

Object shape: We represent the object's shape as an implicit surface $\mathcal{S} \in \mathbb{R}^3$ in the robot's frame, with SDF uncertainty $\mathbf{\Sigma}_\mathcal{S}$ (Refer Section V-C).

Tactile measurements: During interaction, upon detecting contact, we record the corresponding tactile image $\mathbf{I_t}$ and sensor pose $\mathbf{p_t}$:

$$\mathbf{z}_t = \left\{ \mathbf{I}_t \in \mathbb{R}^{640 \times 480 \times 3}, \ \mathbf{p}_t \in SE\left(3\right) \right\} \quad (1)$$

Depth-map: We capture a depth-map $\mathbf{D}_0$ of the object from the camera, represented in the robot-frame: $\mathbf{d}_{1 \cdots M} \in \mathbb{R}^3$.

Assumptions: In line with prior efforts, we assume:
- Calibrated robot-camera extrinsics,
- Fixed object pose and known approximate object dimensions,
- A passive exploration algorithm for object coverage.

The rest of the paper is as follows: Section IV presents a GelSight image to height-map model for tactile perception. Section V combines tactile point-clouds with a depth-map in an incremental GP spatial graph. In Section VI, we demonstrate our method for simulated and real visuo-tactile experiments. Finally, we sum up our efforts in Section VII.

IV. LOCAL SHAPE FROM TOUCH

Vision-based tactile sensors perceive contact geometries as images. The soft, illuminated gelpad deforms elastically on contact and is captured by an embedded camera. We represent local shape recovery as the inverse sensor model:

$$\mathbf{\Omega} : \mathbf{I}_t \mapsto \mathbf{H}_t, \mathbf{C}_t, \text{ where } \begin{array}{l} \mathbf{I}_t\text{: tactile image,} \\ \mathbf{H}_t\text{: recovered height-map,} \\ \mathbf{C}_t\text{: contact area mask} \end{array} \quad (2)$$

With $\mathbf{H}_t$, $\mathbf{C}_t$, and knowledge of sensor pose $\mathbf{p}_t$ from robot kinematics, we can obtain a tactile point-cloud $\mathcal{M}_t$, comprising of 3-D points $\mathbf{m}_i^x$ and normals $\mathbf{m}_i^y$:

$$\mathcal{M}_t = \{[\mathbf{m}_1^x, \mathbf{m}_1^y], [\mathbf{m}_2^x, \mathbf{m}_2^y], \cdots\} \quad (3)$$

In this section, we learn $\mathbf{\Omega}$ through simulation, and its output forms the basis for our visuo-tactile mapping in Section V.
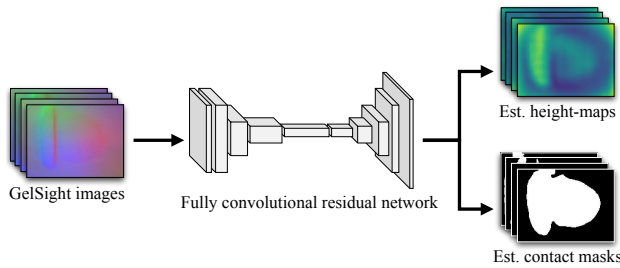
Fig. 2: Our learned model Ω takes in tactile images, and outputs both estimated height-maps and binary contact masks. The residual network is trained on a corpus of GelSight-object interactions in simulation.

Fig. 3: Tactile images generated from GelSight interactions in **[top]** simulated and **[bottom]** real settings. Pictured alongside are the height-maps and contact masks output from our learned model Ω.

A. Learning Ω from simulation

For tactile sensors with soft body deformation, local shape geometry can be learned through supervision. Image-to-depth estimation networks [44, 45] can learn accurate heightmaps from GelSight images. This would require a large corpus of tactile images and corresponding ground-truth depths, for which we can leverage tactile simulation. In particular, Si et al. [23] calibrate their simulator with real-world tactile images, thus mimicking the same intensity distributions.

Network and training: We use an implementation [46] of the fully convolutional residual network [45] as our depth estimator, as shown in Figure 2. The network combines ResNet-50 as the encoder and up-sampling blocks as the decoder. Our model takes tactile images as input, and outputs predictions of both height-map $\mathbf{H}_t$ and contact mask $\mathbf{C}_t$. We choose 30 household objects from YCB dataset [47], and hold out 6 objects for testing generalization. For each object, we generate 660 images from randomly sampled sensor poses on their ground-truth mesh models. We split the train-validation-test sets as 550-50-60.

Benchmarks: We compare Ω with the standard lookup table method [4]. This maps tactile images to gradients of the local shape, and fast Poisson integration generates their height-
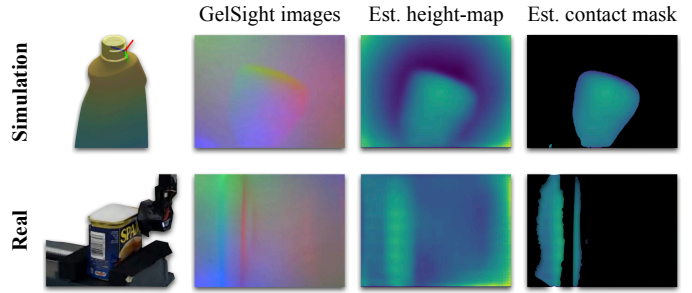
maps. The lookup is built via a calibration routine with a 4 mm sphere indenter. The contact masks are generated by intensity thresholding of contact vs. non-contact frames.

Evaluation: Figure 4 compares Ω with respect to benchmarks on our YCBSight-Sim dataset (refer Section VI-A). We compare each estimated height-map and contact mask against the ground-truth. Specifically, we evaluate:

(i) Pixel-wise RMSE on height-maps, and
(ii) Intersection over union (IoU) on contact masks.

On height-map estimation, we outperform the benchmark with an average RMSE of 0.094 mm across all object classes. The lookup table has larger variance, with an average RMSE of 0.182 mm. Note that the maximum penetration depth of the simulation is 1 mm. On contact mask estimation, we have an average IoU of 0.752, while the handcrafted image thresholding performs worse with 0.379. In addition, the IoU variance appears to be larger for objects with more intricate shapes. Finally, in Figure 3, we show generalization of Ω to both unseen simulation and real-world tactile interactions.

V. 3-D SHAPE ESTIMATION

A. Standard Gaussian processes

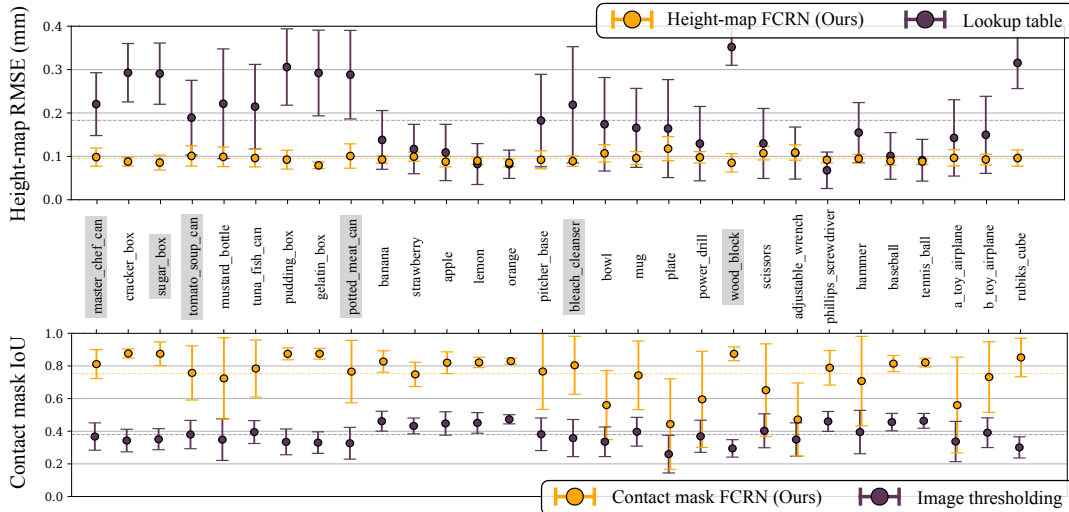A GP is a nonparametric method to learn a continuous function from data, well-suited to model spatial and temporal

Fig. 4: Local shape recovery benchmarked on our YCBSight-Sim dataset (Refer Section VI-A). **[top]** We evaluate our learned model with respect to the baseline lookup table method for height-map estimation. Here we use a pixel-wise root-mean-square error (RMSE)↓ metric, and observe consistent, low error for our method when compared with the lookup table. **[bottom]** We compare our learned contact mask model against intensity-based thresholding on intersection over union (IoU)↑ metric. The test data is randomly generated, and `gray` represents the hold-out objects not encountered in training.

phenomena [48]. To estimate shape, a classical GP considers the object's SDF to be a joint Gaussian distribution over noisy measurements of its surface. At any given point in space, the SDF φ represents the signed-distance from the surface: $\varphi = 0$ on the surface, $\varphi < 0$ inside, and $\varphi > 0$ outside. The GP meaningfully approximates the global shape, even in regions lacking sensor information. Given a dense tactile measurement $\mathbf{m}_i \in \mathcal{M}_t$[1], we learn a function between positions $\mathbf{m}_i^x$ and normals $\mathbf{m}_i^y$:

$$\mathbf{m}_i^x \mapsto [\varphi = 0, \ \mathbf{m}_i^y] \tag{4}$$

More generally, treating the left and right hand side of Equation 4 as the GP's input-output:

$$\mathbf{X} = \{\mathbf{x}_i \in \mathbb{R}^3\}^{1\cdots N} \mapsto \mathbf{Y} = \{\mathbf{y}_i \in \mathbb{R}^4\}^{1\cdots N} \tag{5}$$

The posterior distribution at a query point $(\mathbf{x}_j^*, \mathbf{y}_j^*)$ for a full GP with N measurements, is given by [48]:

$$\mathbf{y}_j^* \sim \mathcal{GP}\big(\underbrace{k_*^T \left(K + \sigma_{\mathrm{n}}^2 I\right)^{-1} Y}_{\text{mean}}, \ \underbrace{k_{**} - k_*^T \left(K + \sigma_{\mathrm{n}}^2 I\right)^{-1} k_*}_{\text{variance}}\big) \tag{6}$$

where σ_{n} is the sensor noise covariance, and $K \in \mathbb{R}^{4N \times 4N}$, $k_* \in \mathbb{R}^{4N \times 4}$ and $k_{**} \in \mathbb{R}^{4 \times 4}$ are the train-train, train-query, and query-query kernels respectively. Each kernel's constituent block $k_{ij} = k(x_i, \ x_j)$ is an $\mathbb{R}^{4 \times 4}$ kernel basis, in our case a thin-plate function [28]. This inference is computationally intractable for the large N that accrues from high-dimensional tactile measurements. The update operations involve costly $O(N^3)$ matrix inversions, and per-query costs $O(N^2)$ (Refer Equation 6). We now present a local approximation that can be updated and queried incrementally, with bounded computational costs.

B. GP-SG: Gaussian process spatial graph

We represent the scene as a spatial factor graph [38], comprising of nodes we optimize for and factors that constrain them. These query nodes $\mathbf{Y}^*$ are at their respective spatial positions $\mathbf{X}^*$, distributed in an S^3 volume. Our optimization goal is to recover the posterior $\hat{\mathbf{Y}}^*$, which represents the SDF of the volume and its underlying uncertainty.

Implementing the full GP (Equation 6) in the graph is costly, as each measurement $(\mathbf{x}_i, \mathbf{y}_i)$ constrains all query nodes $\mathbf{Y}^*$. Motivated by prior work in spatial partitioning [35, 36], we decompose the GP into local unary factors as a sparse approximation. Given that $\mathbf{y}_i$ and query node $\mathbf{y}_j^*$ follow a GP, the joint distribution and conditional are:

$$\begin{bmatrix} \mathbf{y}_i \\ \mathbf{y}_j^* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} k_{ii} + \sigma_{\mathrm{n}}^2 I & k_{i*} \\ k_{*i} & k_{**} \end{bmatrix}\right)$$
$$\mathbf{y}_j^* \mid \mathbf{y}_i \sim \mathcal{N}\left(\underbrace{k_{*i}\left(k_{ii} + \sigma_{\mathrm{n}}^2 I\right)^{-1}\mathbf{y}_i}_{\mu_{\mathbf{y}_j^* \mid \mathbf{y}_i}}, \underbrace{k_{**} - k_{*i}^2 k_{ii}^{-1}}_{\Sigma_{\mathbf{y}_j^* \mid \mathbf{y}_i}}\right) \tag{7}$$

This gives us a unary Gaussian potential which can be incorporated into a least-square setting:

$$\mathcal{G}_{ij} = ||\mathbf{y}_j^* - \mu_{\mathbf{y}_j^* \mid \mathbf{y}_i}||^2_{\Sigma_{\mathbf{y}_j^* \mid \mathbf{y}_i}} \tag{8}$$

[1] or depth map $\mathbf{d}_{1\cdots M} \in \mathbf{D}_0$

Noisy surface readings

Estimated SDF

Estimated uncertainty

Gaussian process spatial graph

● Query grid points
■ Surface measurements
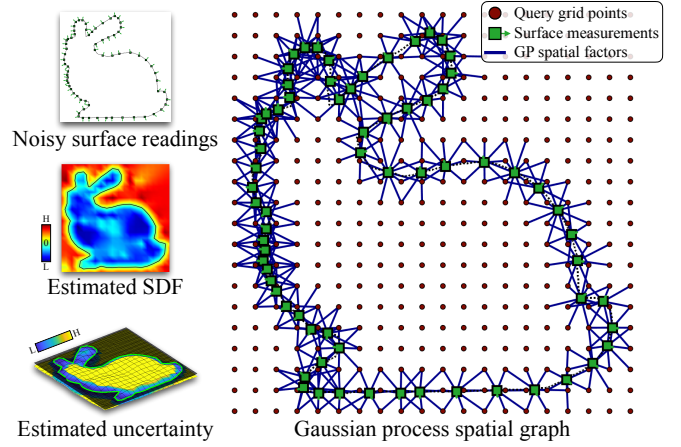— GP spatial factors

Fig. 5: [right] A 2-D illustration of our GP spatial graph (GP-SG), an efficient local approximation to a full GP. The graph consists of SDF query nodes (●) $\mathbf{Y}^*$ each at their spatial positions $\mathbf{X}^*$. Each surface measurement (■) $(\mathbf{x}_i, \mathbf{y}_i)$ produces a unary factor (—) $\mathcal{G}_{ij}$ at query node $\mathbf{y}_j^*$ (within the local radius r). This represents a local Gaussian potential for the GP implicit surface. **[left]** Optimizing for $\hat{\mathbf{Y}}^*$ yields posterior SDF mean + uncertainty. The zero-level set of the SDF gives us the implicit surface $\mathcal{S}$.

At a timestep t, given n measurements $(\mathbf{x}_i, \mathbf{y}_i)$, we add the set of associated factors within a local radius r of each query node's position $\mathbf{x}_j^*$. r represents a tradeoff between speed and reconstruction accuracy, and is empirically set to 15% of the ground-truth object's side length. Thus, for all query nodes, we accumulate a small set of factors:

$$\mathcal{R}_t \triangleq \left\{ \{\mathcal{G}_{ij}\}_{\substack{i=1\cdots n \\ j=1\cdots S^3}} \ \mid \ \|\mathbf{x}_j^* - \mathbf{x}_i\| \le r \right\} \tag{9}$$

This sparsifies an otherwise intractable optimization, pictorially represented in Figure 5 for a 2-D case. Taking the Stanford bunny as an example, we illustrate how a set of noisy surface measurements are converted into local GP factors. The final optimization recovers a posterior SDF mean and uncertainty. More specifically, for the visuo-tactile problem, the *maximum a posteriori* estimation is:

$$\hat{\mathbf{Y}}^* = \underset{\mathbf{Y}^*}{\operatorname{argmin}} \ \overbrace{\sum_{\mathcal{G}_d \in \mathcal{R}_0} \mathcal{G}_d}^{\text{depth factors}} + \overbrace{\sum_{t=1}^{T} \sum_{\mathcal{G}_m \in \mathcal{R}_t} \mathcal{G}_m}^{\text{tactile factors}} + \overbrace{\sum_{\mathbf{y}^* \in \mathbf{Y}^*} ||\mathbf{y}^* - \mathbf{b}||^2_{\Sigma_{\mathbf{b}}}}^{\text{GP priors}} \tag{10}$$

where $\mathcal{R}_0$ is the factor set from the depth-map $\mathbf{D}_0$, and $\mathcal{R}_t$ is the factor set from tactile measurement $\mathcal{M}_t$. The term $\mathbf{b}$ applies a positive SDF prior to nodes, initializing the volume as empty space. Inference is carried out at each timestep via incremental smoothing and mapping (iSAM2) [49].

This framework combines the computational benefits of an online local GPIS [35, 36] with those of an incremental least-squares solver. This is well-suited for sensors like the GelSight, as the dense point-clouds are too expensive to incorporate into a full GP. When querying, we recover the posterior mean and covariance only for the nodes updated— the remaining grid is accessed from cache.

C. Implicit surface generation

The posterior estimate $\hat{\mathbf{Y}}^*$ represents the SDF's mean and uncertainty, sampled from the S^3 volume. A marching cubes algorithm [50] can give us both the implicit surface $\mathcal{S}$ and
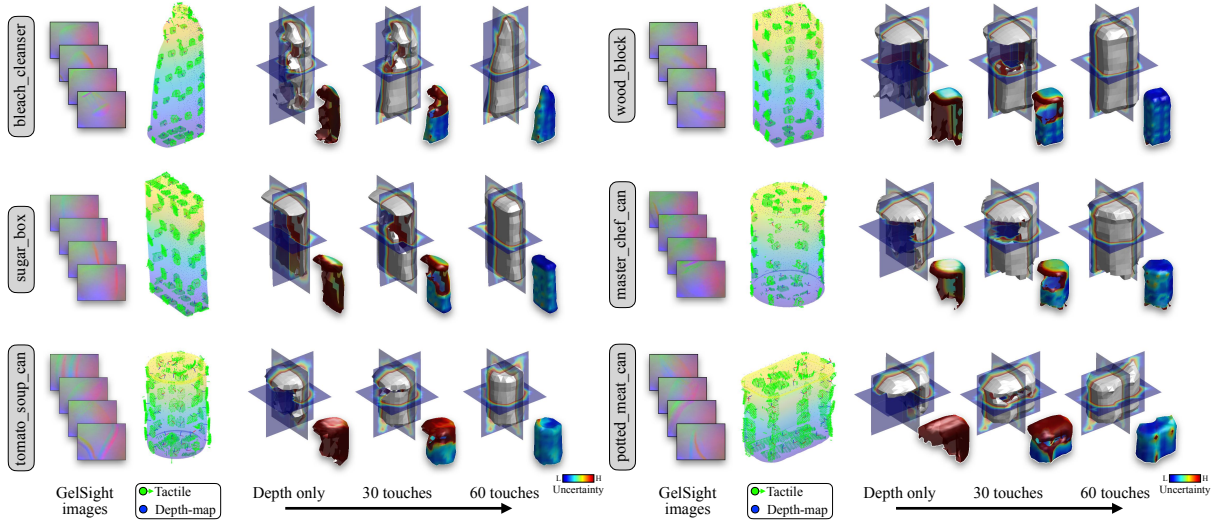
Fig. 6: Results from simulated visuo-tactile mapping on our `YCBSight-Sim` dataset. Shown for each object are (i) sample GelSight images, (ii) tactile and depth-map measurements on the ground-truth mesh, and (iii) frames from the incremental mapping. Each object is initialized with a noisy rendered depth-map (*Depth only*), and with each sequential GelSight measurement, we gain further understanding of global shape and reduce surface uncertainty. Visualized here are the implicit surface + SDF + uncertainty for the intervals of $[0, 30, 60]$ touches.

the corresponding SDF uncertainty $\Sigma_{\mathcal{S}}$. $\mathcal{S}$ is generated as the zero-level set of the SDF:

$$\mathcal{S} \triangleq \{\mathbf{s} \in \mathbb{R}^3 \mid \hat{\mathbf{Y}}^*(\mathbf{s})_{\varphi} = 0\} \quad (11)$$

Finally, we prune faces/vertices from $\mathcal{S}$ that lie outside r for *any* of the sensor measurements. These areas have high surface uncertainty, and our spatial graph will poorly approximate them. Furthermore, this is necessary for sequential data as we cannot expect a watertight mesh from partial coverage.

VI. EXPERIMENTAL EVALUATION

We illustrate our method in simulated (Section VI-B) and real-world (Section VI-C) visuo-tactile experiments. The shape estimates are compared with respect to the ground-truth using the Chamfer distance (CD) [51], a commonly-used shape similarity metric.

Implementation: The framework is executed on an Intel Core i7-7820HQ CPU, 32GB RAM without GPU parallelization. We use the GTSAM [52] optimizer with iSAM2 [49]
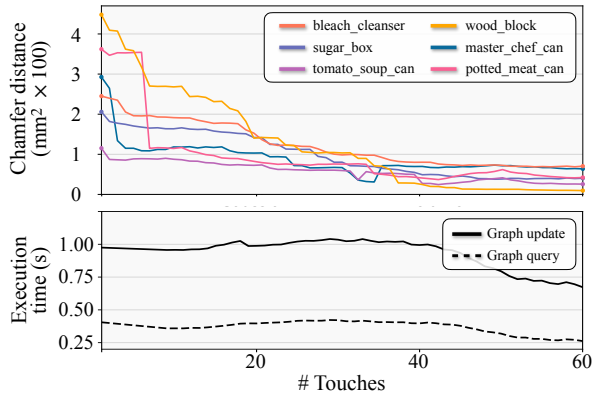

Fig. 7: [top] The Chamfer distance (CD) with respect to ground-truth meshes for `YCBSight-Sim` experiments. Objects are initialized with high CD from partial depth-map, but converge to low-error in 35–40 touches. **[bottom]** Average execution time for update/query operations on our GP spatial graph (GP-SG). At each touch we add $\approx 10^3$ GP factors during update, and recover posterior mean/covariance during query. We see a dip in timing towards the end, due to smaller contact areas on the top of objects.

for incremental inference. Due to the precision of sensing, we empirically weight the noise of tactile measurements lower than that of the depth-map. We set the grid size $S = 16$, which can be increased for higher-resolution reconstructions.

A. Visuo-tactile data collection

We collect the `YCBSight-Sim` and `YCBSight-Real` datasets for evaluating our method. This comprises of YCB ground-truth meshes [47], GelSight images from interaction, sensor poses, and a depth-map. While we consider 30 household objects in simulation, we restrict our shape mapping evaluation to 6 objects. This subset of objects have varied geometries (curved, rectangular, and complex) to verify the generalization of our method.

`YCBSight-Sim`: We generate GelSight-object interactions using *Taxim*, an example-based tactile simulator [23]. We simulate 60 uniformly spread sensor poses on each object, normal to the local surface of the mesh. We render a depth-map from the perspective of an overlooking camera using Pyrender [53]. Finally, zero-mean Gaussian noise is added to tactile point-clouds, sensor poses, and depth-map.

`YCBSight-Real`: We use a UR5e 6-DoF robot arm, mounting the GelSight sensor on a WSG50 parallel gripper. The depth-map is captured via a fixed-pose, calibrated Azure Kinect, approximately 1m away from the object. Our complete setup can be seen in Figure 9. The GelSight captures 640×480 RGB images of the interactions in a 2.66 cm^2 area. The objects are secured by a mechanical bench vise at a known pose, to ensure they remain static. After capturing the depth-map $\mathbf{D}_0$, we approach each object from a discretized set of angles and heights. This simple strategy works well for the selected objects, and future work can replace this with a closed-loop planner. We detect contact events by thresholding the tactile images. We collect 40 tactile images $\{\mathbf{I}_1, \ldots, \mathbf{I}_{40}\}$ of the object's surface (≈ 20 minutes) with the corresponding gripper poses $\{\mathbf{p}_1, \ldots, \mathbf{p}_{40}\}$ via robot kinematics.
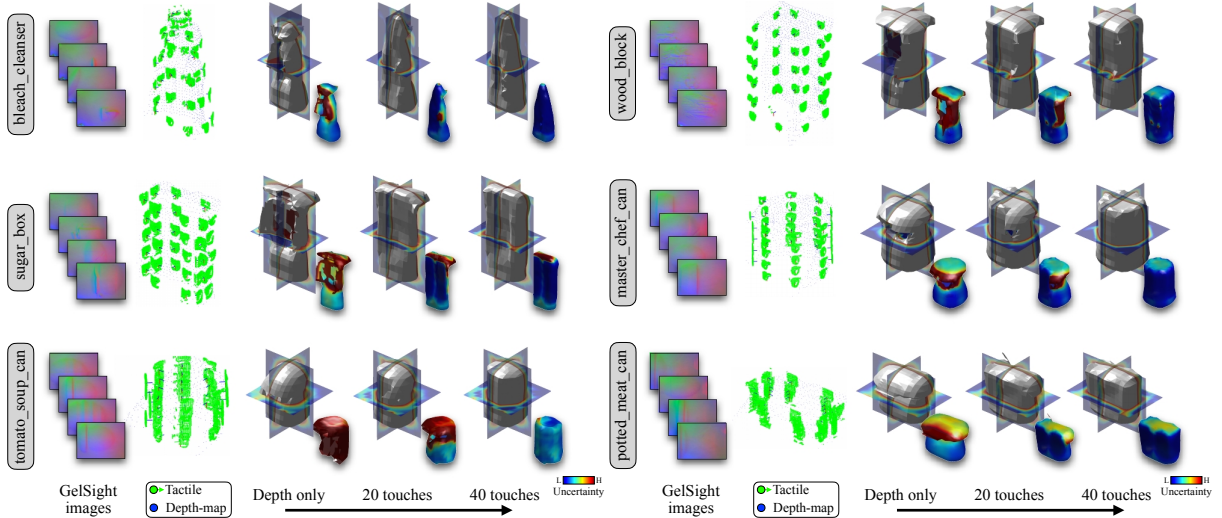
Fig. 8: Results from real visuo-tactile mapping on our `YCBSight-Real` dataset. This is structured similar to Figure 6, except with reconstruction frames at intervals of [0, 20, 40] touches. The Kinect performs poorly for specular objects such as `tomato_soup_can` and `potted_meat_can`, but high-precision GelSight measurements can disambiguate global shape. Our mapping generalizes well and we observe similar results between simulated and real experiments.
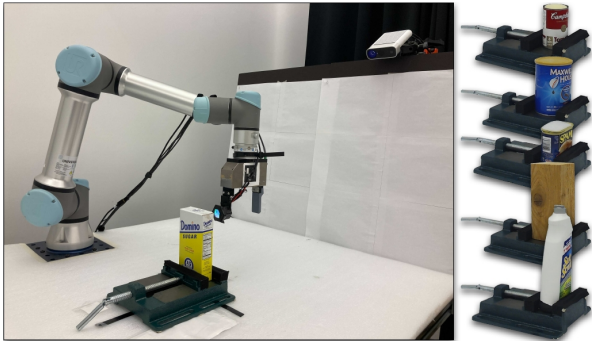

Fig. 9: Experimental setup for the `YCBSight-Real` dataset, with a GelSight tactile sensor, a depth-camera and the YCB objects. Objects are firmly secured on a mechanical bench vise, to ensure they stay stationary. We collect measurements by approaching from a discretized set of angles and heights, and detecting contact from the tactile images. The overlooking Kinect collects a depth-map to initialize our visuo-tactile mapping.
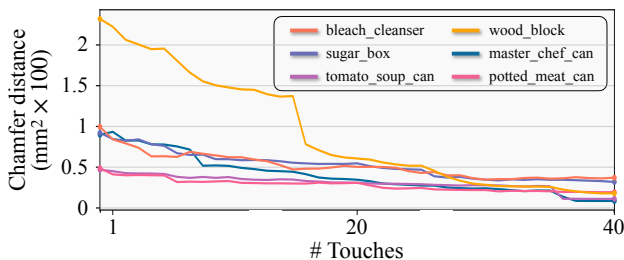

Fig. 10: The Chamfer distance (CD) with respect to ground-truth meshes for our `YCBSight-Real` experiments. We observe the error converges to a similar magnitude as Figure 7 after 30 touches. They initially start out with a lower error than simulation as a result of the hallucinated base measurements we add to each object (refer Section VI-C).

B. Simulated tactile mapping

In Figure 6 we highlight mapping results for the 6 objects in `YCBSight-Sim`. We first visualize the implicit surface and SDF uncertainty from depth-map only. After this, touch measurements are added incrementally and reflect in the shape estimate. The surface uncertainty is typically high for regions that lack depth/tactile information, and reduces over time. Figure 7 shows that the CD with respect to the ground-truth mesh decreases with greater number of touches, and converges within 35–40 touches. The timing plot of graph operations shows near-constant graph update and query time. This reduces towards the end of the datasets due to smaller contact areas on the top surface of the objects. These timings can be further improved by parallelizing spatial operations.

C. Real-world tactile mapping

In Figure 8, we show our method working on real data collected in `YCBSight-Real`. The Kinect depth-maps for specular objects like `tomato_soup_can` and `potted_meat_can` are erroneous, but tactile information provides more precise local shape. To prevent damage to the robot and sensor, we do not explore near the base of the object—we instead hallucinate measurements at the bottom based on the nearest corresponding sensor poses. In Figure 10, we plot the CD over time for the 6 YCB objects. The initial error is lower than simulation due to the additional hallucinated measurements. We see the error converge to an average CD of 18.3 mm^2, a similar magnitude as in the simulated experiments. For reference the average diagonal of the ground-truth YCB objects is 20 cm.

VII. CONCLUSION

We present an incremental framework for 3-D shape estimation from dense touch and vision. We formulate a GP spatial graph (GP-SG) structure, that efficiently infers an object's implicit surface and SDF uncertainty. To integrate GelSight tactile images, we recover local shape with a model learned in tactile simulation. Our method is first demonstrated in a simulated visuo-tactile setting, and is later shown to generalize to real-world shape perception.

As future work, we wish to actively reconstruct these shapes using surface uncertainty information. The current method can further benefit from (i) parallelized spatial graph operations, and (ii) data-driven shape priors [11, 26]. Finally, we wish to consider relaxing the fixed-pose assumption [43], and perception of deformable objects.

REFERENCES

[1] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon, "Kinectfusion: Real-time dense surface mapping and tracking," in *Proc. Int. Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 127–136, IEEE, 2011.

[2] H. B. Helbig and M. O. Ernst, "Optimal integration of shape information from vision and touch," *Experimental brain research*, vol. 179, no. 4, pp. 595–606, 2007.

[3] A. Yamaguchi and C. G. Atkeson, "Combining finger vision and optical tactile sensing: Reducing and handling errors while cutting vegetables," in *Proc. IEEE-RAS Intl. Conf. on Humanoid Robots (Humanoids)*, pp. 1045–1051, IEEE, 2016.

[4] W. Yuan, S. Dong, and E. H. Adelson, "GelSight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, vol. 17, no. 12, p. 2762, 2017.

[5] E. Donlon, S. Dong, M. Liu, J. Li, E. Adelson, and A. Rodriguez, "GelSlim: A high-resolution, compact, robust, and calibrated tactile-sensing finger," in *Proc. IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pp. 1927–1934, IEEE, 2018.

[6] B. Ward-Cherrier, N. Pestell, L. Cramphorn, B. Winstone, M. E. Giannaccini, J. Rossiter, and N. F. Lepora, "The TacTip family: Soft optical tactile sensors with 3D-printed biomimetic morphologies," *Soft robotics*, vol. 5, no. 2, pp. 216–227, 2018.

[7] A. Alspach, K. Hashimoto, N. Kuppuswamy, and R. Tedrake, "Soft-bubble: A highly compliant dense geometry tactile sensor for robot manipulation," in *Proc. IEEE Intl. Conf. on Soft Robotics (RoboSoft)*, pp. 597–604, IEEE, 2019.

[8] M. Lambeta, P.-W. Chou, S. Tian, B. Yang, B. Maloon, V. R. Most, D. Stroud, R. Santos, A. Byagowi, G. Kammerer, *et al.*, "DIGIT: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation," *IEEE Robotics and Automation Letters (RA-L)*, vol. 5, no. 3, pp. 3838–3845, 2020.

[9] A. Padmanabha, F. Ebert, S. Tian, R. Calandra, C. Finn, and S. Levine, "OmniTact: A multi-directional high-resolution touch sensor," in *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pp. 618–624, IEEE, 2020.

[10] S. Wang, Y. She, B. Romero, and E. Adelson, "GelSight Wedge: Measuring high-resolution 3D contact geometry with a compact robot finger," in *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, IEEE, 2021.

[11] S. Wang, J. Wu, X. Sun, W. Yuan, W. T. Freeman, J. B. Tenenbaum, and E. H. Adelson, "3D shape perception from monocular vision, touch, and shape priors," in *Proc. IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pp. 1606–1613, IEEE, 2018.

[12] M. Bauza, O. Canal, and A. Rodriguez, "Tactile mapping and localization from high-resolution tactile imprints," in *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pp. 3811–3817, IEEE, 2019.

[13] E. J. Smith, R. Calandra, A. Romero, G. Gkioxari, D. Meger, J. Malik, and M. Drozdzal, "3D shape reconstruction from vision and touch," in *Proc. Conf. on Neural Information Processing Systems (NeurIPS)*, 2020.

[14] E. J. Smith, D. Meger, L. Pineda, R. Calandra, J. Malik, A. Romero, and M. Drozdzal, "Active 3D shape reconstruction from vision and touch," in *Proc. Conf. on Neural Information Processing Systems (NeurIPS)*, 2021.

[15] A. Hertzmann and S. M. Seitz, "Example-based photometric stereo: Shape reconstruction with general, varying BRDFs," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 27, no. 8, pp. 1254–1264, 2005.

[16] M. K. Johnson and E. H. Adelson, "Retrographic sensing for the measurement of surface texture and shape," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 1070–1077, 2009.

[17] M. K. Johnson, F. Cole, A. Raj, and E. H. Adelson, "Microgeometry capture using an elastomeric sensor," *ACM Trans. Graph.*, vol. 30, July 2011.

[18] P. Sodhi, M. Kaess, M. Mukadam, and S. Anderson, "Patchgraph: In-hand tactile tracking with learned surface normals," in *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2022.

[19] R. Ambrus, V. Guizilini, N. Kuppuswamy, A. Beaulieu, A. Gaidon, and A. Alspach, "Monocular depth estimation for soft visuotactile sensors," in *Proc. IEEE Intl. Conf. on Soft Robotics (RoboSoft)*, 2021.

[20] M. Bauza, E. Valls, B. Lim, T. Sechopoulos, and A. Rodriguez, "Tactile object pose estimation from the first touch with geometric contact rendering," in *Proc. Conf. on Robot Learning, CoRL*, 2020.

[21] A. Agarwal, T. Man, and W. Yuan, "Simulation of vision-based tactile sensors using physics based rendering," in *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pp. 1–7, IEEE, 2021.

[22] S. Wang, M. M. Lambeta, P.-W. Chou, and R. Calandra, "TACTO: A fast, flexible, and open-source simulator for high-resolution vision-based tactile sensors," *IEEE Robotics and Automation Letters (RA-L)*, 2022.

[23] Z. Si and W. Yuan, "Taxim: An example-based simulation model for gelsight tactile sensors," *IEEE Robotics and Automation Letters (RA-L)*, 2022.

[24] M. Björkman, Y. Bekiroglu, V. Högman, and D. Kragic, "Enhancing visual perception of shape through tactile glances," in *Proc. IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pp. 3180–3186, IEEE, 2013.

[25] J. Ilonen, J. Bohg, and V. Kyrki, "Three-dimensional object reconstruction of symmetric objects by fusing visual and tactile sensing," *Intl. J. of Robotics Research (IJRR)*, vol. 33, no. 2, pp. 321–341, 2014.

[26] J. Varley, C. DeChant, A. Richardson, J. Ruales, and P. Allen, "Shape completion enabled robotic grasping," in *Proc. IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pp. 2442–2447, IEEE, 2017.

[27] G. Z. Gandler, C. H. Ek, M. Björkman, R. Stolkin, and Y. Bekiroglu, "Object shape estimation and modeling, based on sparse Gaussian process implicit surfaces, combining visual data and tactile exploration," *J. of Robotics and Autonomous Systems (RAS)*, vol. 126, p. 103433, 2020.

[28] O. Williams and A. Fitzgibbon, "Gaussian process implicit surfaces," *Gaussian Proc. in Practice*, pp. 1–4, 2007.

[29] A. Bierbaum, I. Gubarev, and R. Dillmann, "Robust shape recovery for sparse contact location and normal data from haptic exploration," in *Proc. IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pp. 3200–3205, IEEE, 2008.

[30] S. Dragiev, M. Toussaint, and M. Gienger, "Gaussian process implicit surfaces for shape estimation and grasping," in *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pp. 2845–2850, IEEE, 2011.

[31] S. Ottenhaus, M. Miller, D. Schiebener, N. Vahrenkamp, and T. Asfour, "Local implicit surface estimation for haptic exploration," in *Proc. IEEE-RAS Intl. Conf. on Humanoid Robots (Humanoids)*, pp. 850–856, IEEE, 2016.

[32] N. Jamali, C. Ciliberto, L. Rosasco, and L. Natale, "Active perception: Building objects' models using tactile exploration," in *Proc. IEEE-RAS Intl. Conf. on Humanoid Robots (Humanoids)*, pp. 179–185, IEEE, 2016.

[33] Z. Yi, R. Calandra, F. Veiga, H. van Hoof, T. Hermans, Y. Zhang, and J. Peters, "Active tactile object exploration with Gaussian processes," in *Proc. IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pp. 4925–4930, IEEE, 2016.

[34] D. Driess, P. Englert, and M. Toussaint, "Active learning with query paths for tactile object shape exploration," in *Proc. IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pp. 65–72, IEEE, 2017.

[35] B. Lee, C. Zhang, Z. Huang, and D. D. Lee, "Online continuous mapping using Gaussian process implicit surfaces," in *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pp. 6884–6890, IEEE, 2019.

[36] J. A. Stork and T. Stoyanov, "Ensemble of sparse Gaussian process experts for implicit surface mapping with streaming data," in *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pp. 10758–10764, IEEE, 2020.

[37] A. Ranganathan, M.-H. Yang, and J. Ho, "Online sparse Gaussian process regression and its applications," *IEEE Trans. on Image Processing*, vol. 20, no. 2, pp. 391–404, 2010.

[38] F. Dellaert and M. Kaess, "Factor graphs for robot perception," *Foundations and Trends in Robotics*, vol. 6, no. 1-2, pp. 1–139, 2017.

[39] X. Yan, V. Indelman, and B. Boots, "Incremental sparse GP regression for continuous-time trajectory estimation and mapping," *J. of Robotics and Autonomous Systems (RAS)*, vol. 87, pp. 120–132, 2017.

[40] D. M. Rosen, G. Huang, and J. J. Leonard, "Inference over heterogeneous finite-/infinite-dimensional systems using factor graphs and Gaussian processes," in *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pp. 1261–1268, IEEE, 2014.

[41] M. Mukadam, J. Dong, X. Yan, F. Dellaert, and B. Boots, "Continuous-time Gaussian process motion planning via probabilistic inference," *Intl. J. of Robotics Research (IJRR)*, vol. 37, no. 11, pp. 1319–1340, 2018.

[42] J. Wang, T. Shan, and B. Englot, "Underwater terrain reconstruction

from forward-looking sonar imagery," in *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pp. 3471–3477, IEEE, 2019.

[43] S. Suresh, M. Bauza, K.-T. Yu, J. G. Mangelson, A. Rodriguez, and M. Kaess, "Tactile SLAM: Real-time inference of shape and pose from planar pushing," in *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, May 2021.

[44] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in Neural Information Processing Systems*, vol. 3, pp. 2366–2374, 2014.

[45] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *Proc. Intl. Conf. on 3D Vision (3DV)*, pp. 239–248, IEEE, 2016.

[46] "FCRN: Fully convolutional residual network for depth estimation." https://github.com/XPFly1989/FCRN, 2018.

[47] B. Calli, A. Singh, J. Bruce, A. Walsman, K. Konolige, S. Srinivasa, P. Abbeel, and A. M. Dollar, "Yale-CMU-Berkeley dataset for robotic manipulation research," *Intl. J. of Robotics Research (IJRR)*, vol. 36, no. 3, pp. 261–268, 2017.

[48] C. E. Rasmussen, "Gaussian processes in machine learning," in *Summer School on Machine Learning*, pp. 63–71, Springer, 2003.

[49] M. Kaess, H. Johannsson, R. Roberts, V. Ila, J. Leonard, and F. Dellaert, "iSAM2: Incremental smoothing and mapping using the Bayes tree," *Intl. J. of Robotics Research (IJRR)*, vol. 31, no. 2, pp. 216–235, 2012.

[50] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3D surface construction algorithm," *ACM SIGGRAPH Computer Graphics*, vol. 21, no. 4, pp. 163–169, 1987.

[51] H. G. Barrow, J. M. Tenenbaum, R. C. Bolles, and H. C. Wolf, "Parametric correspondence and chamfer matching: Two new techniques for image matching," tech. rep., 1977.

[52] F. Dellaert, "Factor graphs and GTSAM: A hands-on introduction," tech. rep., Georgia Institute of Technology, 2012.

[53] M. Matl, "Pyrender." https://github.com/mmatl/pyrender, 2019.