Contact-Aware Probabilistic Reconstruction for Contact-Rich Manipulation

***Abstract*— Robotic assembly tasks such as peg-in-hole insertion require precise geometric reasoning, yet visual sensing noise in real-world systems often exceeds the tight tolerances required for successful insertion. In this work, we propose *ContactFusion*, a probabilistic mapping framework that fuses depth sensing and force–torque measurements to estimate the geometry of insertion targets. Our method builds a Stochastic Poisson Surface Map (SPSMap), an uncertainty-aware implicit surface representation constructed using Stochastic Poisson Surface Reconstruction (SPSR). To incorporate contact information, we introduce a sampling based contact location estimator that converts force–torque measurements into spatial hypotheses over candidate contact locations on the robot end-effector. These hypotheses are fused with depth observations within a sequential reconstruction framework, enabling the map to be refined through both visual and contact interactions. We evaluate ContactFusion in simulation and on a real robotic system in a peg-in-hole setting. Our results show that SPSMap produces more accurate and geometrically consistent reconstructions, improving reconstruction F-score by up to 30–35%, while providing uncertainty estimates that enable active reconstruction strategies.**

I. Introduction

Dexterous manipulation often requires precise geometric reasoning; however, sensor noise in real robot applications often exceeds the necessary tolerances. Tasks such as part alignment, constrained sliding, snap-fitting, and peg-in-hole insertion demand accurate estimates of the relative pose between interacting objects. As tolerances tighten, millimeter-scale geometric errors can lead to jamming, excessive contact forces, or failure to complete the task. This makes geometric state estimation a crucial component for the success of current manipulation systems.

Current manipulation pipelines determine the pose of the insertion target by geometric registration of RGB-D measurements against a reference model (e.g., CAD model) [1], or using deep learning-based prediction methods [2]. While vision-based solutions can estimate the target's pose up to the millimeter-level, they can be heavily affected by sensor noise, environmental and robot-induced occlusions, producing misalignment. Exploiting passive mechanical compliance and robust control strategies, such as impedance or compliant control, can reduce contact forces and compensate for this misalignment, at the cost of introducing a tradeoff between compliance and trajectory tracking [3]. One way to tackle such a requirement for precision could be to complement global visual information with fine local contact information. Motivated by recent works [4], we explore this problem from the perspective of building multi-modal scene representations.

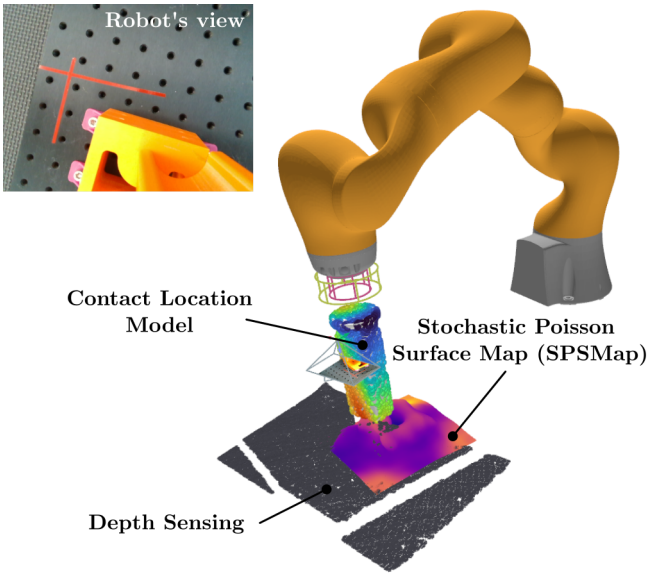

Fig. 1: **ContactFusion builds probabilistic maps of an insertion target.** By building upon the Stochastic Poisson Surface Reconstruction method [5] in a sequential formulation and a new contact location model fused with depth observations, ContactFusion produces continuous estimates of the target's surface and uncertainty that are up to 30% more accurate than prior approaches.

In this work, we introduce *ContactFusion*, a probabilistic mapping framework that builds a geometric representation of the insertion target by fusing depth and contact measurements. Our framework integrates these different sensor modalities using the Stochastic Poisson Surface Reconstruction (SPSR) method [5], producing maps that explicitly model the surface of the target object and its uncertainty. We experimentally demonstrate that by combining vision and contact into this probabilistic representation, we can improve estimates of the target's shape by 30% compared to existing methods, uncertainty estimates of the geometry, and also enable further directions such as active perception.

Our specific contributions are:

- We present a probabilistic mapping framework — *ContactFusion*, which combines vision and contact sensing using the Stochastic Poisson Surface Reconstruction method, into an uncertainty-aware representation— *SPSMap*.
- We introduce a sampling based contact location estimator to transform force/torque measurements into spatial likelihoods over candidate surface contact locations.

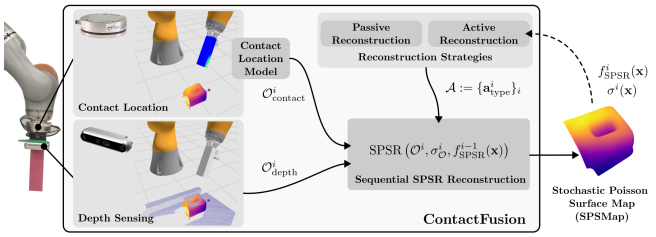

Fig. 2: **ContactFusion system overview.** Depth measurements are fused with Force-Torque derived spatial likelihoods into an uncertainty aware geometric representation. We propose using the Stochasitic Poisson Surface Reconstruction [5] algorithm to sequentially fuse depth and F/T derived contact locations into a probabilistic map.

II. Method

Fig. 2 shows the main steps of the ContactFusion system. The inputs to our system are depth measurements from an RGB-D sensor, as well as force-torque measurements from a sensor located at the robot's wrist. Its output is a Stochastic Surface Reconstruction Map (SPSMap), a field that implicitly encodes the geometry and uncertainty of the map.

A. Probabilistic Surface Mapping

ContactFusion implements a sequential mapping procedure by integrating visual and contact measurements in the form of oriented point sets $\mathcal{O}$. By relying on the SPSR method, it implements online reconstruction by exploiting the operator defined by Eq. (**??**). This motivates the following sequential update rule:

$$f^i_{\text{SPSR}}(\mathbf{x}), \sigma^i(\mathbf{x}) = \text{SPSR}\left(\mathcal{O}^i, \sigma^i_{\mathcal{O}}, f^{i-1}_{\text{SPSR}}(\mathbf{x})\right), \quad (1)$$

where i indexes the measurements, and f^{i-1}_{SPSR} is the estimated scalar field obtained in the last iteration.

This mapping procedure is implicit, since SPSR defines the map via the field f_{SPSR}. Therefore, it does not provide direct access to occupancy or the surface itself. However, since it defines a probability distribution of the map, it enables closed-form geometric queries.

The occupancy probability at a query point $\mathbf{x}$ is defined by the cumulative distribution up to the zero-level set:

$$p\big(f_{\text{SPSR}}(\mathbf{x}) \leq 0\big) = \text{CDF}_{f_{\text{SPSR}}(\mathbf{x})}(0),$$

and the probability that $\mathbf{x}$ lies on the surface is given by the zero-level set:

$$p\big(f_{\text{SPSR}}(\mathbf{x}) = 0\big) = \text{PDF}_{f_{\text{SPSR}}(\mathbf{x})}(0).$$

Fig. **??** illustrates the field that induces the surface as the zero level set (a), its uncertainty (b) and the estimated occupancy (c). We note that the SPSR method yields smooth surface estimates while retaining a probabilistic interpretation of occupancy through the closed-form queries of the implicit function defined above. This makes it well-suited to settings in which local measurements (such as contact) must refine a globally-coherent surface estimate from other modalities, like vision. The next sections provide details about the measurements used for each modality.

B. Vision-based Measurements

A depth observation $\mathcal{O}^i_{\text{depth}}$ is obtained from a depth sensor measurement (e.g, RGB-D camera) and converted into an oriented point set:

$$\mathcal{O}^i_{\text{depth}} := \{(\mathbf{p}_s, \mathbf{n}_s)\}_s, \quad (2)$$

where $\mathbf{p}_s \in \mathbb{R}^3$ are 3D points extracted from the depth measurement. $\mathbf{n}_s \in \mathbb{R}^3$ are corresponding surface normals estimated from local geometry, using the K-Nearest Neighbours method implemented in Open3D [6]. We assume a fixed isotropic variance given by σ^i_{depth}.

C. Contact Location Measurements

While the depth sensor directly outputs the depth observations, a force-torque measurement does not directly yield a contact point. Instead, it constrains the set of possible contact locations consistent with the measured wrench. For this, we introduce a *contact location model* to map force-torque measurement to an oriented point set of contact hypotheses.

Given a force-torque measurement $\mathcal{F}^i = (\mathbf{f}^i_{\text{obs}}, \boldsymbol{\tau}^i_{\text{obs}})$ expressed in the sensor frame, the estimator computes candidate contact locations on the peg surface that are consistent with rigid-body mechanics. Let $\mathbf{p}_{\text{can}}$ denote a candidate point on the peg surface and $\mathbf{p}_{\text{origin}}$ a point at the origin of the sensor frame. Under a single-point contact assumption, the measured torque at the contact point should satisfy

$$\boldsymbol{\tau}^i_{\text{can}} \approx (\mathbf{p}_{\text{origin}} - \mathbf{p}_{\text{can}}) \times \mathbf{f}^i_{\text{obs}}. \quad (3)$$

We therefore define a residual function

$$l(\mathbf{p}_{\text{can}}; \mathcal{F}^i) = \left\| (\mathbf{p}_{\text{origin}} - \mathbf{p}_{\text{can}}) \times \mathbf{f}^i_{\text{obs}} - \boldsymbol{\tau}^i_{\text{obs}} \right\|^2, \quad (4)$$

which measures the consistency between the candidate contact location $\mathbf{p}_{\text{can}}$ and the measured wrench $\boldsymbol{\tau}_{\text{obs}}$.

To construct the contact observation $\mathcal{O}^i_{\text{contact}}$, we collect K observations on the peg surface over a time-horizon, and retain those whose residual falls below a predefined threshold. This procedure is equivalent to a rejection sampling approach: samples inconsistent with the measured wrench are rejected, while those satisfying the residual constraint are accepted as plausible contact hypotheses. The associated contact normal for observation k is taken to align with the measured normal force direction $\mathbf{f}_n$,

$$\mathbf{n}^k = \frac{\mathbf{f}^k_n}{\|\mathbf{f}^k_n\|}. \quad (5)$$

The resulting set of accepted candidates forms the oriented point set:

$$\mathcal{O}^i_{\text{contact}} := \{(\mathbf{p}^k_{\text{can}}, \mathbf{n}^k)\}_{k \in K}. \quad (6)$$

Optionally, a friction cone constraint may be imposed to further eliminate geometrically infeasible contact hypotheses. Similarly to the depth observations, we also consider a fixed isotropic variance $\sigma^i_{\text{contact}}$. The complete procedure and real examples of the contact location model are shown in Fig. 3.

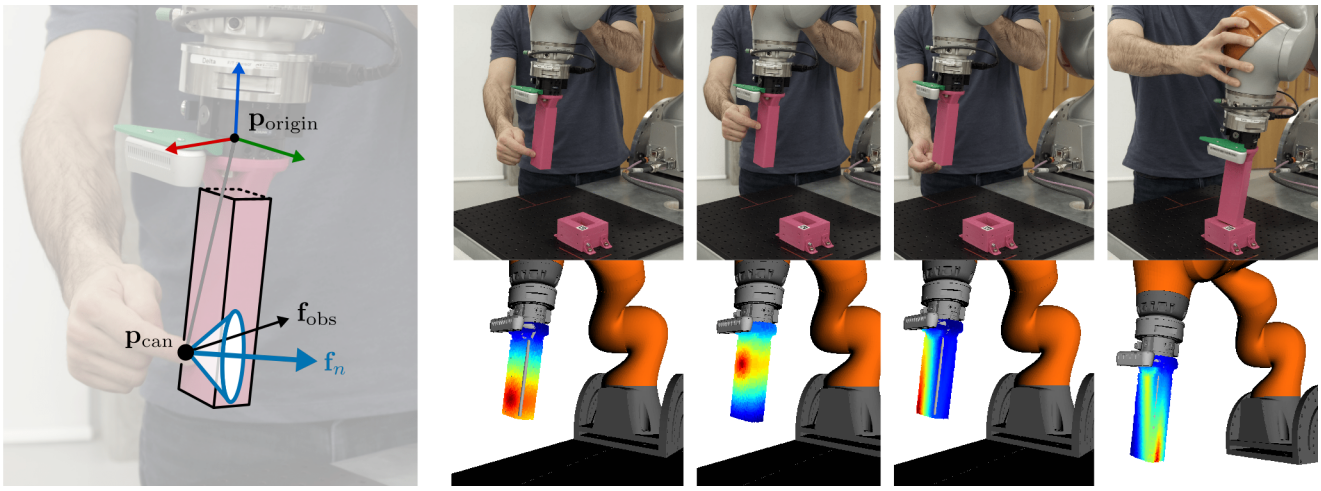

Fig. 3: **Proposed contact location estimator.** Left: Force-torque residual model. $\mathbf{p}_O$ represents sensor origin, $\mathbf{p}_C$ represents sample where likelihood is being computed, $\mathbf{f}_O$ represents translational element of the wrench. Right: Examples of the contact location sensor from real interactions, where the color denotes the likelihood (blue: lowest, red: highest).

D. Visuo-Tactile Reconstruction Strategies

In contrast to vision or LiDAR-based reconstruction approaches, a visuo-tactile system needs to explicitly establish contact with the environment to acquire new measurements. ContactFusion implements two reconstruction strategies, which determine how vision and contact observations are obtained and integrated. Specifically, they determine a sequence of actions indexed by i:

$$\mathcal{A} := \{\mathbf{a}^i_{\text{type}}\}_i, \tag{7}$$

where $\mathbf{a}^i_{\text{type}}$ can be vision action $\mathbf{a}^i_{\text{vision}}$ or a probing action $\mathbf{a}^i_{\text{probing}}$. We assume that each action is followed by an observation.

A visual action specifies a camera viewpoint relative to a planar center in the end-effector frame,

$$\mathbf{a}^i_{\text{vision}} := (x_v,\, y_v,\, \alpha,\, \beta), \tag{8}$$

where (x_v, y_v) defines the center of the viewing configuration in the xy-plane, and α and β represent elevation and azimuth angles defining a single keyframe viewpoint. This defines a hemisphere of rays the robot can align to.

A probing action specifies a planar target location and a commanded translation along the negative z-axis of the end-effector frame,

$$\mathbf{a}^i_{\text{probing}} := (x_p,\, y_p,\, z_p), \tag{9}$$

where (x_i, y_i) parameterizes the probe location in the xy-plane and z_p is the magnitude of the translation along $-z$. This defines a 2D grid of probing locations.

These actions are executed using linear interpolated motion plans tracked via a compliant impedance controller.

a) Passive Reconstruction: The first strategy uses a predetermined sequence of actions for the mapping process. We assume a fixed budget of N measurements for this process. The first γ correspond to vision actions, while the remaining $N-\gamma$ correspond to probing actions. Both vision and probing actions are sampled randomly, given some pre-defined limits to the parameters that define each space.

b) Active Reconstruction: The second strategy exploits the uncertainty estimate of the surface to guide the exploration, effectively implementing a *next best action* approach. Similarly to the passive strategy, we assign a budget of N measurements for the experiment with γ vision actions. However, instead of sampling, we use a greedy strategy guided by estimated uncertainty of the map at i.

For vision actions, we choose the ray in the vision action space that intersects the area of the mesh with higher uncertainty. Similarly, for probing actions we project the uncertainty to the 2D grid and vertically probe the map point with higher value.

III. Experiments and Results

A. Exp. 1: Insertion Target Reconstruction

Protocol: We evaluate reconstruction accuracy in simulation using a fixed measurement budget of $N = 9$, with $\gamma = 3$ depth observations followed by 6 contact probes. Each method is evaluated over 10 randomized object pose scenarios, using identical action sequences across methods. Fig. 4 reports precision and recall metrics averaged over 10 sample scenario at a error threshold $\tau = 2$mm across 9 measurements. Our experimental findings indicate that SPSMap and GPIS recover higher precision and recall when compared to OccMap. We find the SPSMap and GPIS representations extrapolate to unobserved geometry as is evident with the higher recall metrics across both the Rectangle and Arch geometries. Contact measurements improve local geometric conditioning, as can be observed in the increased trends within both precision and recall after measurement 4 on both geometries. Both SPSMap and GPIS provide alternative belief representations over geometry for the insertion task that are able to integrate contact and depth information.

Fig. 5 summarizes reconstruction performance via F-score at $\tau = \{1\text{mm}, 2\text{mm}\}$, averaged across all scenarios and plot-

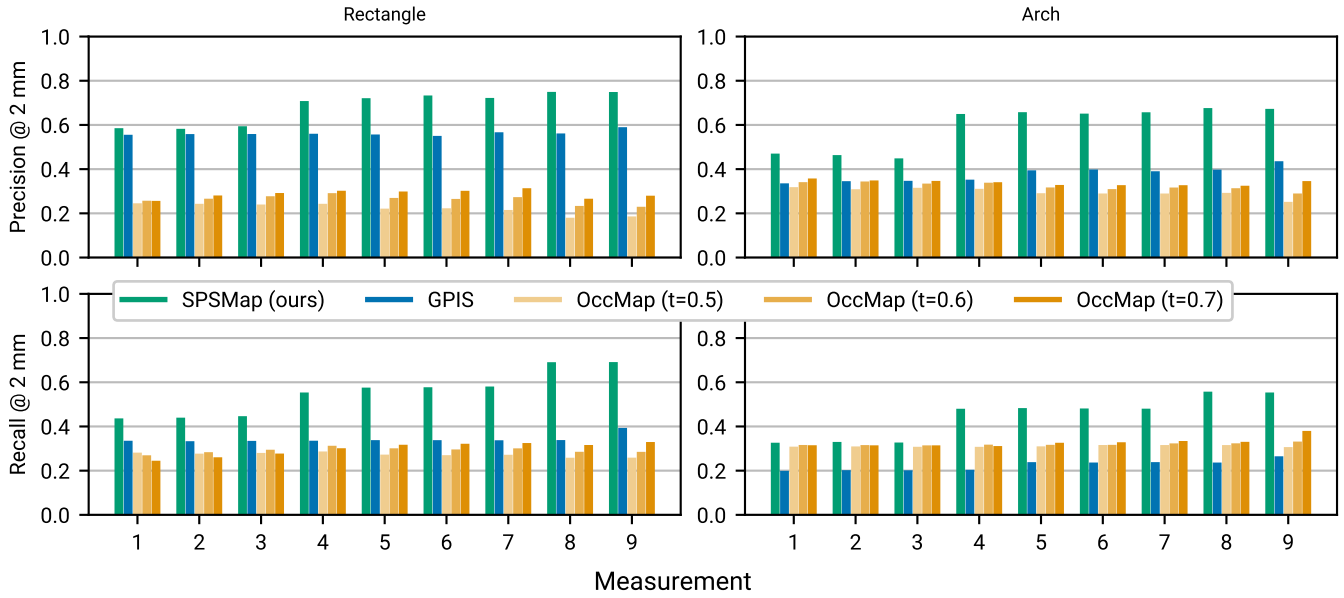

Fig. 4: **Exp. 1: Insertion Target Reconstruction.** We report the Precision and Recall metrics averaged over 10 sample scenarios for the ARCH and RECTANGLE geometry at a $\tau = 2$mm. Large precision indicates a high proportion of the reconstructed mesh is under this τ. Large recall indicates a high proportion of the ground truth mesh under τ. Measurements 1-3 are depth based measurements acquired from pre-scripted robot panning motion. Measurements 3-9 are contact location measurements acquired from pre-scripted robot probe motions. We report surface reconstruction for OccMap at 3 occupancy thresholds, t = $\{0.5, 0.6, 0.7\}$

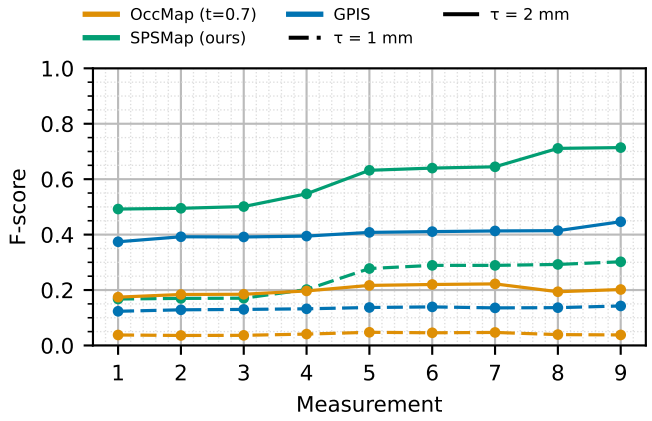

Fig. 5: **Exp. 1: Avg F-scores at different tolerances.** We report the F-score averaged across 10 scenarios, as a function of discrete depth and contact measurements. Measurements 1-3 are depth measurements, measurements 4-9 are contact probes.

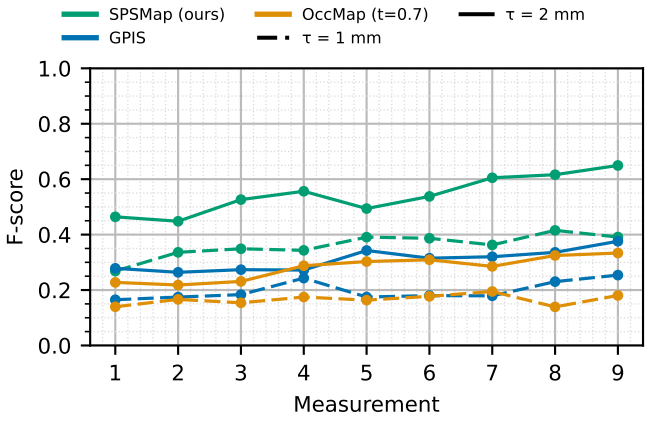

Fig. 6: **Exp. 2: Real Robot Validation.** Avg F-scores at different thresholds, τ, averaged across 5 real robot scenarios. Measurements 1-3 are via depth, measurements 4-9 are contact probes.

ted over measurement index. SPSMap consistently achieves higher F-scores throughout the reconstruction process, with improvements of up to 30%.

B. Exp. 2: Real Robot Validation

For the real robot experiment, Fig. 6 shows the average F-Score for the 5 sequences collected over randomized hole positions. We report that SPSMap has a higher F-Score of up to 35%. Furthermore, for both Exp. 1 and Exp. 2 we observe that SPSMap consistently recovers better precision and recall on symmetric and assymetric objects over GPIS. We hypothesise that the improved performance of SPSMap over GPIS arises from the inductive bias imposed by the positive definite kernels used in GPIS. These kernels enforce symmetric covariance structures in the implicit function, which can bias extrapolation in sparsely observed regions, whereas the SPSR-based formulation underlying SPSMap imposes weaker symmetry assumptions and can therefore better capture asymmetric or partially observed geometries.

REFERENCES

[1] R. Haralick, H. Joo, C. Lee, X. Zhuang, V. Vaidya, and M. Kim, "Pose estimation from corresponding point data," *IEEE Trans. Syst. Man Cybern.*, vol. 19, no. 6, pp. 1426–1446, 1989.

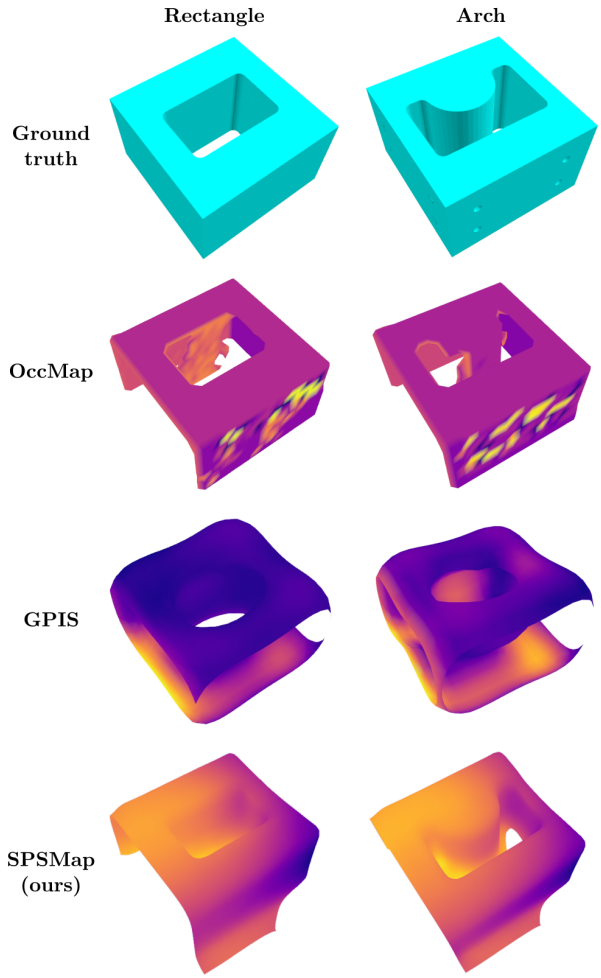

Fig. 7: **Exp. 1: Qualitative results.** We show examples of the different baselines methods for the arch and rectangle geometries.

[2] B. Wen, W. Yang, J. Kautz, and S. Birchfield, "FoundationPose: Unified 6D Pose Estimation and Tracking of Novel Objects," in *IEEE Int. Conf. Computer Vision and Pattern Recognition*, Jun. 2024, pp. 17 868–17 879.

[3] Y. Jiang, Z. Huang, B. Yang, and W. Yang, "A review of robotic assembly strategies for the full operation procedure: Planning, execution and evaluation," *Robotics and Computer-Integrated Manufacturing*, vol. 78, p. 102 366, 2022.

[4] S. Kim and A. Rodriguez, "Active Extrinsic Contact Sensing: Application to General Peg-in-Hole Insertion," in *IEEE Int. Conf. Robot. Autom.*, 2022, pp. 10 241–10 247.

[5] S. Sellán and A. Jacobson, "Stochastic Poisson Surface Reconstruction," *ACM Transactions on Graphics*, vol. 41, no. 6, pp. 1–12, 2022.

[6] Q.-Y. Zhou, J. Park, and V. Koltun, "Open3D: A modern library for 3D data processing," *arXiv:1801.09847*, 2018.