Towards Generating a Probabilistic Ensemble of 3D Rigid Body Simulations from an RGBD Image Using Gaussian Processes

Onur Beker [*1] **Yunhan Wang** [*2] **Barbu Bojor** [*3]

Abstract

The standard approach to generating robot behavior is to optimize the control inputs of a deterministic rigid body simulation. While this approach is remarkably successful in controlled environments like factories, it does not account for uncertainty and is therefore poorly suited for uncontrolled environments. One way to address this limitation is to optimize the controls to maximize the total probability of success across a diverse ensemble of rigid body simulations. As a first step towards building such a simulation framework, we propose a method to build a probabilistic ensemble of 3D reconstructions of an object given a single RGBD image of it. Our method first constructs a dataset of 3D points that are likely to lie within the object boundaries by retrieving geometric primitives from a category-level mesh template and realigning them with the RGBD image. It then fits a Gaussian process to this data and samples from its posterior. We evaluate our method on a real-world cereal box, and show that it can successfully generate an ensemble of likely 3D reconstructions. Our code is available on GitHub.

1. Introduction

Robotic automation of tasks such as agriculture, construction, and elderly care can significantly improve the overall human condition by facilitating access to food, shelter, and dignity for everyone. Such tasks all require a capacity to physically manipulate rigid bodies. The state of the art for solving this problem involves: i) constructing a 3D rigid body simulation (i.e., a first order ODE derived from combining Newton–Euler equations (Lanczos, 1949) with various contact models (Mason, 2001)) where all physical parameters are *precisely specified* (e.g., geometry, inertia, coefficients of friction), ii) optimizing control inputs to this ODE to minimize time and energy expenditure. This succeeds when all physical parameters can be *precisely measured* (e.g., factories), and fails in uncontrolled environments (e.g., fields, construction sites, private dwellings).

In contrast, consider how humans solve rigid body manipulation problems. Imagine a person looking at their coffee cup. From the image they know where the front surface points are, but they do not exactly know how the occluded 3D geometry continues behind or how much the coffee cup weighs. Yet it is trivial for them to expect a roughly cylindrical surface 5-10 cm behind and a mass between 100-1000 grams, and attempt a grasp that *remains sensible for most realizations of physical parameters* within these intervals.

This illustrates how human behavior is driven not by precise descriptions of what is known, but by sensible expectations about what is not known. Forming such expectations requires a framework to express particular degrees of being imprecise (e.g., a roughly cylindrical surface 5-10 cm behind), in other words a calibrated notion of uncertainty. To account for such considerations, we propose constructing an ensemble of rigid body simulations with different parameters each (i.e., rather than a single simulation with precisely specified parameters), where every simulation in the ensemble tracks its own probability of being the true state of the world. Control inputs can then be optimized to maximize the *total probability of success* across the entire ensemble.

The focus of this *data literacy* project is to create a subcomponent of such a simulation framework, by implementing a *data collection and analysis pipeline* to build a probabilistic ensemble for the complete 3D geometry of an unknown object from a single RGBD image of it (where only the surfaces facing the camera are visible and the 3D geometry in occluded regions are unknown). Section 2 describes the two main steps of our pipeline: i) creating a dataset of 3D point coordinates and associated labels that capture whether the coordinate lies inside or outside the object, ii) fitting a Gaussian process (GP) (MacKay, 1998; Rasmussen et al., 2006; Hennig et al., 2022) to this data and sampling from the resulting posterior to build an ensemble of likely 3D reconstructions (i.e., including estimates of the occluded surfaces). Sec.3 demonstrates an application of this data collection and analysis pipeline on a real-world object (a cereal box) for which a ground truth geometric model was obtained from a database of photogrammetry scans.[1]

[*]Equal contribution [1]Matrikelnummer 6601173, onur.beker@student.uni-tuebingen.de, MSc Machine Learning [2]Matrikelnummer 6606475, yunhan.wang@student.uni-tuebingen.de, MSc Machine Learning [3]Matrikelnummer 6609773, barbu.bojor@student.uni-tuebingen.de, MSc Machine Learning.

Project report for the "Data Literacy" course at the University of Tübingen, Winter 2023/24 (Module ML4201). Style template based on the ICML style files 2023.

[1]https://poly.cam/

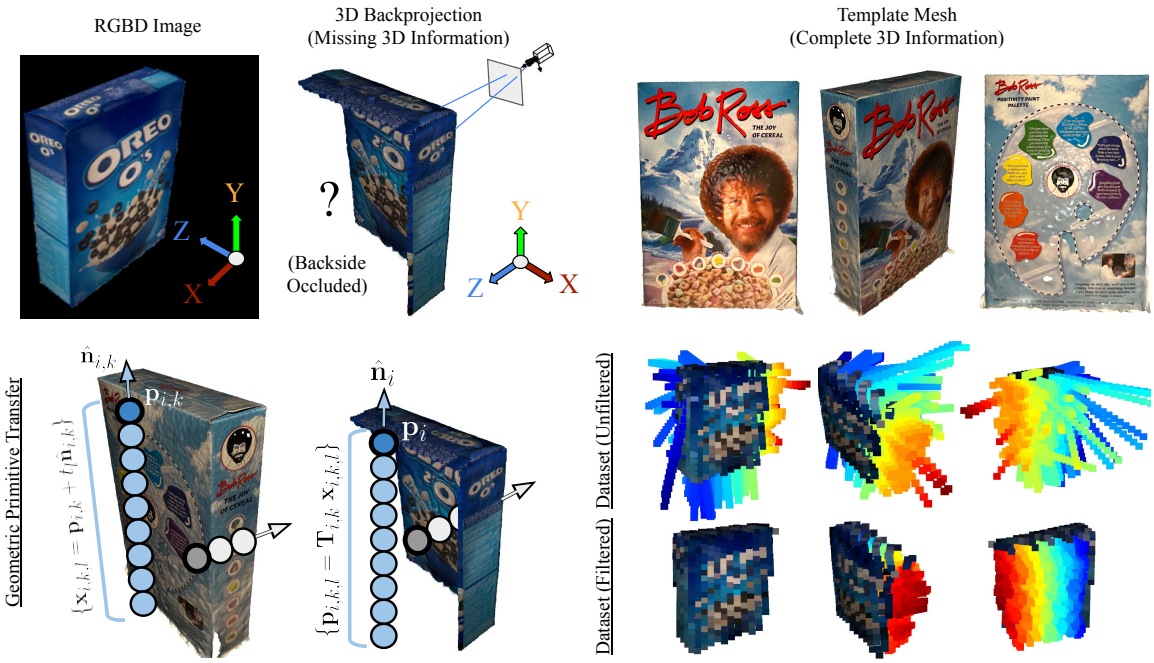

Figure 1. Given an RGBD image I of an object $\mathcal{O}$ (i.e., Oreo cereal box) and a 3D complete template mesh $\mathcal{T}$ (i.e., Bob Ross cereal box), we create a dataset of 3D points covering the occluded regions in I by transferring geometric primitives (i.e., points on the lines along surface normals) from $\mathcal{T}$ to $\mathcal{O}$ based on visually similar points between them. We then process this dataset by filtering outliers.

2. Methods and Data

Preprocessing: Given an RGBD image of an object $\mathcal{O}$ (i.e., Oreo cereal box in Fig.1), a set of pixel-wise descriptors $\{\mathbf{z}_i \in \mathbb{R}^{1024}\}_{i=1}^{900}$ is extracted using the 30×30 spatial value tokens of the final attention layer from the DINOv2 ViT backbone (Oquab et al., 2023). These descriptors are backprojected into 3D using depth values, and per point surface normals (i.e., unit vectors perpendicular to the object surface) are computed, resulting in a descriptor pointcloud $\mathcal{S}_\mathcal{O} = \{(\mathbf{p}_i, \hat{\mathbf{n}}_i, \mathbf{z}_i)\}_{i=1}^{900}$ where every 3D point $\mathbf{p}_i \in \mathbb{R}^3$ has an associated descriptor $\mathbf{z}_i$ and surface normal $\hat{\mathbf{n}}_i \in \mathbb{R}^3$. We also assume a template mesh $\mathcal{T}$ from the same semantic category as $\mathcal{O}$ (i.e., Bob Ross cereal box in Fig.1) is available with complete 3D information, processed in the same way to create a descriptor pointcloud $\mathcal{S}_\mathcal{T} = \{(\mathbf{p}_j, \hat{\mathbf{n}}_j, \mathbf{z}_j)\}_{j=1}^{3002}$.

Creating the Dataset: The essential property of the descriptors is that the inner product $\mathbf{z}_i^T \mathbf{z}_j$ captures visual similarity. This similarity metric is utilized to transfer geometric primitives from the 3D complete template $\mathcal{S}_\mathcal{T}$ to the occluded regions of $\mathcal{S}_\mathcal{O}$ through the following process:

- For each $\mathbf{p}_i \in \mathcal{S}_\mathcal{O}$, its top-K nearest-neighbors (NN) $\{\mathbf{p}_{i,k}\}_{k=1}^{10} \subset \mathcal{S}_\mathcal{T}$ are found according to $\mathbf{z}_i^T \mathbf{z}_{i,k}$.
- For every NN point $\mathbf{p}_{i,k} \in \mathcal{S}_\mathcal{T}$, an evenly spaced set of 50 points along the line $\{\mathbf{x}_{i,k,l} = \mathbf{p}_{i,k} + t_l \hat{\mathbf{n}}_{i,k}\}_{l=1}^{50}$ are found, and for each $\mathbf{x}_{i,k,l}$, its signed distance (SDF) $d_{i,k,l} = \min_{\mathbf{y} \in \partial\mathcal{S}_\mathcal{T}} ||\mathbf{y} - \mathbf{x}_{i,k,l}||$ is computed.
- All points and SDF values $\{(\mathbf{x}_{i,k,l}, d_{i,k,l})\}$ are currently expressed in the coordinate frame of $\mathcal{S}_\mathcal{T}$. To transfer them to the coordinate frame of $\mathcal{S}_\mathcal{O}$, the 3D coordinate transformations $\mathbf{T}_{i,k} \in SE(3)$ that align the lines $\mathbf{p}_i + t\hat{\mathbf{n}}_i$ and $\mathbf{p}_{i,k} + t\hat{\mathbf{n}}_{i,k}$ are computed.
- The dataset $\mathcal{D} = \{(\mathbf{p}_{i,k,l}, d_{i,k,l})\}_{i=1,k=1,l=1}^{900\ 10\ 50}$ is created through the transformation $\mathbf{p}_{i,k,l} = \mathbf{T}_{i,k}\, \mathbf{x}_{i,k,l}$.

Processing the Dataset: Delving into the dataset and creating a quick 3D plot of the entries in $\mathcal{D} = \{(\mathbf{p}_{i,k,l}, d_{i,k,l})\}$, we identify two main sources of noise: i) imperfect estimation of normals $\hat{\mathbf{n}}_i$ around surface discontinuities (e.g., edges and corners of the cereal box) where there is no unique definition of a surface normal, ii) the discrepancy between the sizes of the imaged object $\mathcal{O}$ and the template mesh $\mathcal{T}$ (i.e., the Bob Ross cereal box is much larger). To filter these effects, we observe that: i) presence of surface discontinuities and imperfect surface normals correlate with a high total curvature $\kappa(\mathbf{p}_i)$ around a point $\mathbf{p}_i \in \mathcal{S}_\mathcal{O}$, ii) if a point $\mathbf{p}_{i,k,l} \in \mathcal{D}$ is retrieved from a region of $\mathcal{T}$ with significant size discrepancy compared to $\mathcal{O}$ (e.g., from a longer edge), it will protrude out compared to other points in $\mathcal{D}$ and will hence have a larger average distance $N(\mathbf{p}_{i,k,l})$ to its nearest N neighbors. We therefore remove any outlier points in $\mathcal{D}$ that have a curvature larger than a fixed threshold $\kappa((p)) \geq 0.05$, or that satisfies $|N(\mathbf{p}_{i,k,l}) - \mu_N| \geq \frac{\sigma_N}{2}$ where (μ_N, σ_N) denote the mean and standard deviation of the metric $N(\mathbf{p}_{i,k,l})$ across $\mathcal{D}$.

Fitting a GP: The goal of this step is to create an estimator for the confidence that any 3D coordinate $\mathbf{v} \in \mathbb{R}^3$ (poten-

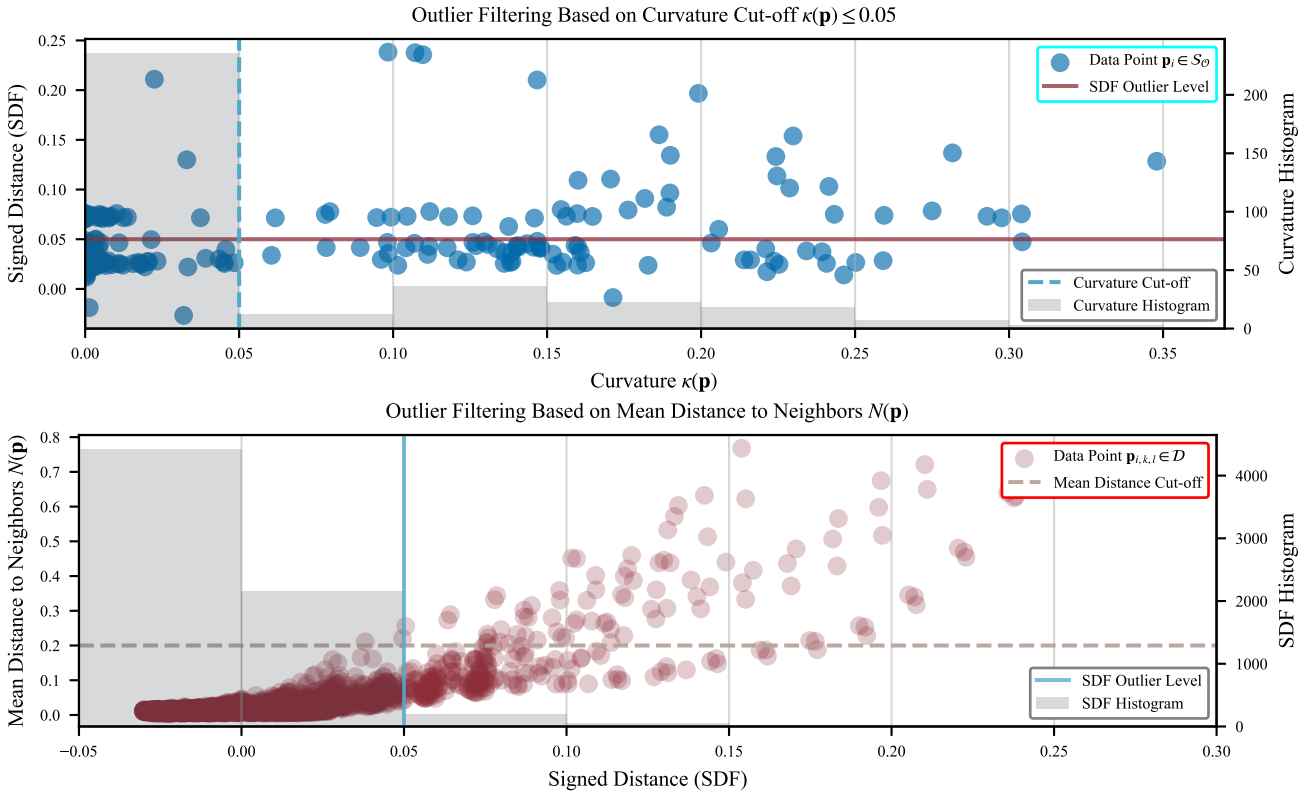

Figure 2. We analyze our dataset to find two suitable metrics for filtering outliers: i) total curvature $\kappa(\mathbf{p})$ around a point, ii) mean distance $N(\mathbf{p})$ between a point and its nearest N neighbors. Plots above show that these surrogate metrics correlate highly with the ground truth (GT) signed distance metric. Dashed lines show suitable cut-off values based on the surrogate metrics (i.e., our filter), while the solid lines show suitable cut-off values based on the GT metric (i.e., best possible filter assuming access to GT). Histogram values show that both filters preserve most of the data points and discard only relatively few outliers (i.e., filtering does not annihilate the dataset).

tially from occluded regions in the RGBD image) lies inside the object $\mathcal{O}$. We will then sample from this estimator to create our ensemble. We construct this estimator as follows:

- We create a 3D grid of points $\{\mathbf{v}_n\}$ that are 1 cm apart from each other along all xyz axes.
- For every point $\mathbf{v}_n$, we count the number c_n of data points in $\mathcal{D}$ that satisfy $||\mathbf{v}_n - \mathbf{p}_{i,k,l}|| \leq d_{i,k,l}$.
- We normalize the counts c_n to lie between $[0, 1]$, and fit a GP to $\mathbf{X} = \{\mathbf{v}_n \in \mathbb{R}^3\}, \mathbf{Y} = \{c_n \in \mathbb{R}\}$.

The GP model uses an RBF kernel with a fixed length scale of 1 cm to match the spatial resolution of the grid $\{\mathbf{v}_n\}$. The variance of the Gaussian likelihood is in turn optimized using L-BFGS (Liu & Nocedal, 1989) to maximize the log-likelihood of generating the data $(\mathbf{X}, \mathbf{Y})$. For all of these operations, we use the GPy library. [2]

3. Results

Processing the dataset: As mentioned in Sec.2, data is processed through two filtering steps to remove outliers: i) removing any point $\mathbf{p}_{i,k,l}$ whose retrieval involves a point $\mathbf{p}_i \in \mathcal{S}_\mathcal{O}$ with curvature $\kappa(\mathbf{p}_i) \geq 0.05$, ii) removing any point $\mathbf{p}_{i,k,l}$ whose average Euclidean distance $N(\mathbf{p}_{i,k,l})$ to its nearest 30 neighbors in $\mathcal{D}$ is half a standard deviation or further away from the mean.

Fig.2 visualizes the filtering results by highlighting the correlations between the filtering metrics $\kappa(\mathbf{p}_i)$ and $N(\mathbf{p}_{i,k,l})$ and the ground truth signed distance (SDF) metric (i.e., a correctly retrieved point would lie inside the object and hence have an SDF value of zero or less). We note that the SDF values in the plots are measured using the ground truth 3D model for the Oreo cereal box and they *cannot* be measured from the single RGBD image that is the input for our method. This is why we are searching for surrogate metrics like $\kappa(\mathbf{p}_i)$ and $N(\mathbf{p}_{i,k,l})$ that we *can* measure from an RGBD image which correlate with SDF values, so when we remove points with a high surrogate metric it is likely that we are also removing outlier points with high SDF values. Looking at the plots, we can reach two main conclusions:

- The solid lines denote the ground truth SDF value beyond which we would call a point an outlier, and dashed lines denote the cut-off value of the surrogate metric beyond which points are filtered out (i.e., surro-

[2] https://github.com/SheffieldML/GPy

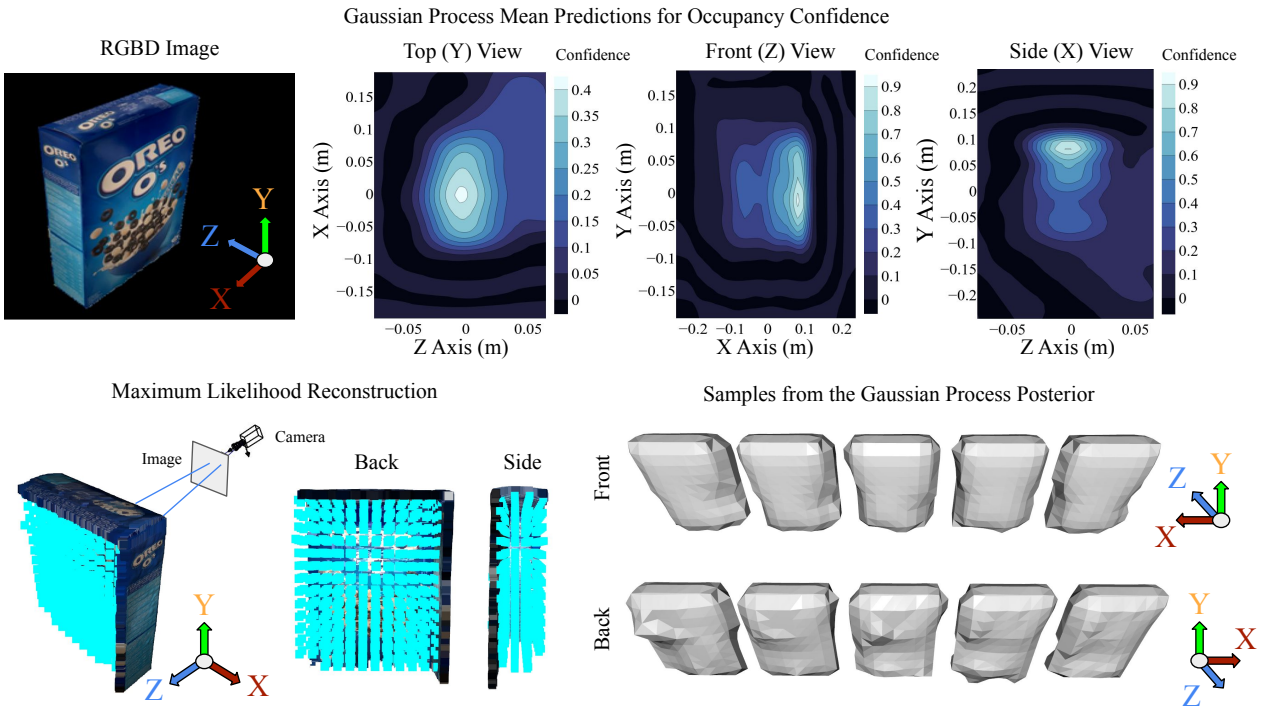

Figure 3. We convert our dataset to occupancy confidences by counting the number of data points that cover each 3D coordinate on a uniform grid and normalizing the result between $[0, 1]$. Finally, we fit a Gaussian process (GP) model and sample from its posterior to obtain an ensemble of likely 3D reconstructions. Plots above show cross-sections of the GP mean (whose domain is $\mathbb{R}^3$) along the XZ, XY, and YZ planes. The 3D meshes for the maximum likelihood reconstruction and posterior samples are obtained from thresholding the GP mean and posterior values respectively, and applying the marching cubes algorithm.

gate outliers). We can see that the corresponding two regions have a large intersection region (i.e., the upper-right quadrants in both plots), meaning the surrogate metrics strongly correlate with the ground truth SDF values and are therefore suitable for filtering.

- Looking at the histograms, we can see that most of the data points lie below the cut-off values, meaning there are relatively few outliers and filtering does not remove too many data points (i.e., sufficiently many data points remain to fit a GP).

Fitting a GP: After the data is filtered, it is converted to occupancy confidences as described in Sec.2, which are then used to fit a GP. Fig.3 shows 2D slices of the mean of the resulting GP measure over $\mathbb{R}^3$, the associated maximum-likelihood reconstruction, and 5 samples from the GP posterior to create the ensemble. The main conclusions are:

- The GP mean captures a sensible representation of the occupancy confidences and the resulting maximum likelihood reconstruction is well aligned with the ground truth 3D model for the Oreo cereal box.
- Samples from the GP capture a proper representation of the uncertainty in 3D geometry. In particular, it can be seen that the samples have more variance on the back surface compared to the front, which is expected since the back surface is occluded and has higher uncertainty.

4. Discussion & Conclusion

Given a single RGBD image of an object, our pipeline successfully builds a probabilistic ensemble of likely 3D reconstructions that appropriately represents the uncertainty associated with regions that are not directly visible (i.e., occluded) in the image. Looking at Fig.3 makes one limitation of our method clear: it does not have any inductive biases to model correlations in how certain surface points tend to deform together. For example, we know that cereal boxes are likely to be symmetric and have planar surfaces, but no part of our method actually communicates this information to the GP, so samples from the posterior exhibit local deformations that resemble spikes rather than deformations that preserve symmetry and coplanarity. To address this, future work can fit the GP model not directly on individual 3D coordinates, but on the parameters of more complex geometric primitives such as superquadrics (Barr, 1981).

Contribution Statement

All members actively participated in all aspects of the project. Onur Beker led the overall project, Yunhan Wang focused on processing the dataset and filtering visualizations, and Barbu Bojor focused on fitting the GP model and 3D visualizations of the posterior samples.

References

Barr, A. H. Superquadrics and angle-preserving transformations. *IEEE Computer graphics and Applications*, 1(1): 11–23, 1981.

Hennig, P., Osborne, M. A., and Kersting, H. P. *Probabilistic numerics: computation as machine learning*. Cambridge University Press, 2022.

Lanczos, C. *The variational principles of mechanics*. University of Toronto Press, 1949.

Liu, D. C. and Nocedal, J. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

MacKay, D. Introduction to gaussian processes. *NATO ASI Ser. F Comput. Syst. Sci.*, 168:133, 1998.

Mason, M. T. *Mechanics of robotic manipulation*. 2001.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Rasmussen, C. E., Williams, C. K., et al. *Gaussian processes for machine learning*, volume 1. Springer, 2006.