Robotics: Science and Systems 2024
Delft, Netherlands, July 15-July 19, 2024

You've Got to Feel It To Believe It: Multi-Modal Bayesian Inference for Semantic and Property Prediction

Parker Ewen, Hao Chen, Yuzhen Chen, Anran Li, Anup Bagali, Gitesh Gunjal, and Ram Vasudevan
University of Michigan, Ann Arbor, USA

***Abstract*—Robots must be able to understand their surroundings to perform complex tasks in challenging environments and many of these complex tasks require estimates of physical properties such as friction or weight. Estimating such properties using learning is challenging due to the large amounts of labelled data required for training and the difficulty of updating these learned models online at run time. To overcome these challenges, this paper introduces a novel, multi-modal approach for representing semantic predictions and physical property estimates jointly in a probabilistic manner. By using conjugate pairs, the proposed method enables closed-form Bayesian updates given visual and tactile measurements without requiring additional training data. The efficacy of the proposed algorithm is demonstrated through several simulation and hardware experiments. In particular, this paper illustrates that by conditioning semantic classifications on physical properties, the proposed method quantitatively outperforms state-of-the-art semantic classification methods that rely on vision alone. To further illustrate its utility, the proposed method is used in several applications including to represent affordance-based properties probabilistically and a challenging terrain traversal task using a legged robot. In the latter task, the proposed method represents the coefficient of friction of the terrain probabilistically, which enables the use of an on-line risk-aware planner that switches the legged robot from a dynamic gait to a static, stable gait when the expected value of the coefficient of friction falls below a given threshold. Videos of these case studies as well as the open-source C++ and ROS interface can be found at https://roahmlab.github.io/multimodal_mapping/**

I. Introduction

Scene understanding from exteroceptive sensing via images or point clouds enables mobile robots to perform object avoidance, terrain traversal, and a variety of other tasks. The introduction of high-level and task-dependent semantic labels for scene understanding has recently spurred rapid growth in this area [1], [2]. To extract semantic labels from images or point clouds, it is common to employ semantic segmentation neural networks; however, even state-of-the-art methods return inconsistent labels under viewpoint or lighting changes, and are heavily reliant on the data used for training. Unfortunately, collecting data to train these networks is expensive and out-of-distribution examples may lead to incorrect classifications.

To address the challenge of viewpoint inconsistency, several state-of-the-art methods project image-based semantic classifications onto metric maps [3], [4]. This provides a shared frame of reference between images thereby enabling the application of probabilistic methods to fuse semantic data. While such methods demonstrate improved performance over state-of-the-art image-based segmentation [5], these methods have been restricted to either visual or lidar data. This is limiting since the properties of a semantic class are often of primary interest rather than the class labels themselves while completing a robotic task. For instance, the coefficient of friction is more informative than the terrain class for a legged robot. While recent semantic mapping methods have derived recursive filters for image-based semantic predictions, properties conditioned on these semantic classes are often assumed to be known *a priori* and are considered immutable at run time [3], [5], [6].

This paper introduces a multi-modal semantic prediction and property estimation framework to jointly estimate se-

Fig. 1: The method proposed in this paper jointly estimates semantic classifications and physical properties by combining visual and tactile data into a single semantic mapping framework. RGB-D images are used to build a metric-semantic map that iteratively estimates semantic labels. A property measurement is taken which in turn updates both the semantic class predictions and physical property estimates. In the depicted example, the robot is unsure if the terrain in front of it is snow or ice from vision measurements alone (prior estimates) which dramatically affects the coefficient of friction and the associated gait that can be applied to safely traverse the terrain. The robot uses a tactile sensor attached to its manipulator to update its coefficient of friction estimation (posterior estimates), which then enables it to change gaits to cross the ice safely.

This work is supported by the Ford Motor Company and the National Science Foundation Career Award #1751093.

mantic labels alongside physical properties as depicted in Figure 1. The key insight is that a semantic class may be conditioned on its physical properties and physical properties may likewise be conditioned on the semantic class. In this way, vision-based semantic classifications can be used to build a measurement model for physical properties, and tactile-based sensing can be used to update the semantic predictions via property measurements. To apply this insight, this paper demonstrates how to recursively estimate physical properties from semantic class predictions using only visual information. Subsequently, this paper describes how to exploit a tactile sensing modality to update the semantic classification and property estimates probabilistically and in closed-form using Bayesian inference.

The efficacy of the proposed approach is demonstrated on several hardware platforms including a robotic manipulator and a legged robot. First, we show that by leveraging friction measurements obtained through tactile sensing we can improve semantic classification performance when compared to state-of-the-art semantic mapping methods that rely solely on visual information. Second, we demonstrate that our approach estimates physical properties more accurately than a method using vision alone. Multiple sensing modalities are used to estimate the coefficient of friction of the terrain surrounding a robot, which enables the use of an online risk-aware planner that switches the legged robot from a dynamic gait to a static, stable gait if the expected value of the coefficient of friction of the terrain falls below a given threshold. Importantly, the proposed multi-modal framework allows the legged robotic system to complete a challenging terrain traversal task whereas state-of-the-art vision-based traversability estimation methods fail to accurately predict the traversability of the terrain resulting in the robot slipping and falling. Finally, we show the broad applicability of the proposed framework for affordance-based property prediction by demonstrating door opening with unknown push-pull interactions.

The contributions of this work are two-fold: first, a technique to apply conjugate prior theory to enable the joint filtering of data collected from visual and tactile sensing modalities for semantic classification and property estimation; and second, an approximate conjugate prior for the Gaussian Mixture distribution that enables computationally tractable filtering. This enables closed-form Bayesian updates given visual semantic classifications and tactile property measurements. We demonstrate that by leveraging this probabilistic approach, we are able to improve semantic prediction performance across multiple metrics in simulation using only one property measurement. These results are validated in hardware demonstrations. Two case studies are subsequently presented that illustrate the utility of performing this type of multi-modal property estimation. An open-source C++ and ROS package for the robot-agnostic implementation of the proposed approach as well as videos of the case studies are provided on the project page website[1]. To the best of our knowledge, this is the first application of conjugate priors for fusing visual and tactile sensing modalities for semantic classification and property estimation.

The remainder of this paper is organized as follows: Section II summarizes the semantic prediction and property estimation literature and Section III reviews preliminary material used throughout the paper. Section IV provides an overview of our method. Section V introduces the vision-based semantic prediction pipeline and Section VI the tactile-based semantic prediction and property estimation pipeline. Section VII describes the implementation. Sections VIII and IX describe the evaluation of our algorithm in simulation and using hardware demonstrations, respectively. Section X discusses future work and provides concluding remarks.

II. Related Works

This section describes related work in semantic prediction and property estimation as well as the gap in the literature this work addresses.

A. Semantic Prediction

State-of-the-art methods for semantic classification almost exclusively rely on neural networks to produce pixel-wise semantic predictions [1], [7], [8]. Often these approaches use task-dependent object designations as semantic labels [9], [10] such as abstract topological information [11] or material classifications [12]. Unfortunately, semantic prediction methods often have no temporal consistency, meaning two subsequent images in a video stream may have vastly different semantic predictions.

To overcome this challenge and relate semantic predictions spatially and temporally, early work in semantic mapping projected semantic predictions onto a geometric representation forming a metric-semantic map [13]. In this way, image-based semantic predictions have a shared frame of reference within this map. These early methods employed voting-like schemes, regularization, and other algorithmic methods to fuse semantic predictions [4], [14].

Recent work in semantic mapping treats the semantic fusion process probabilistically to recursively estimate semantic labels [3], [15], building off of prior probabilistic approaches to binary occupancy mapping [16]. These methods treat pixel-wise semantic predictions as measurements and probabilistically fuse these measurements over the map using Bayesian inference. By approaching semantic prediction in this way, these approaches are robust to noisy pixel-wise semantic classifications and have been shown to be more accurate than methods that rely on image-based classifications alone [5].

Multi-modal semantic prediction, and the corresponding semantic mapping extension, have been around since the rise of modern computer vision [17]. Much of this work comes from the vision community and, as such, many of these methods rely on different vision-based modalities such as depth [18], thermal imaging [19], and scale-invariant feature transform (SIFT) factors [20]. Many modern multi-modal semantic classification methods are learning-based, where the

[1] https://roahmlab.github.io/multimodal_mapping/

multi-modal visual features are combined either at the input of or downstream within neural networks to output semantic classifications [21], [22].

B. Property Estimation

While semantic classification methods have seen widespread use in the robotics community, there are many robotic tasks (e.g., footstep planning, grasp planning, manipulation) where the underlying physical properties of the semantic classes are critical. Additionally, if semantic classifications are incorrect, the resultant property estimates may also be incorrect, leading to task failure.

Previous work in property estimation has focused on estimating the expected value for the coefficient of friction from vision alone [23]–[25]. These methods have been extended in a recent paper that estimates probability distributions for the coefficient of friction given visual semantic class estimates by conditioning properties on semantic classes [5].

Other methods for property estimation focus specifically on estimating traversability [3], [26], [27]. This is done to bypass the need to estimate terrain properties by instead estimating the regions in the environment a robot can traverse using visual features [28]. This is often an ill-posed problem as traversability is not only terrain-dependent, but also dependent on the motion of the robot (e.g., walking vs. running) [29]. For example, it is possible for small insects to walk across a large body of water, but if a large robot applied the same gait it would most likely sink.

Notably, there has been a recent shift towards multi-modal property estimation in the literature. One such approach proposes an online, self-supervised learning framework to estimate traversability by back-projecting previously traversed regions into the camera frame and training a network in a self-supervised manner to correctly label these previously seen regions as traversable [30]. This method is considered multi-modal because it incorporates an implicit goal-success signal, in the form of the robot having not fallen over, alongside visual information. Other work aims to encode property features in a latent space using tactile and visual information to train reinforcement learning agents [31]. Recently, [32] proposed a method for learning motion policies by training a network to relate image-based features and physical properties such as friction in simulation. It was shown that the trained network accurately predicted properties in hardware demonstrations, however these predictions only output a single scalar value for properties without considering uncertainty in the prediction. A similar framework for multi-modal probabilistic property estimation within a semantic mapping framework was also demonstrated for the purposes of traversability estimation [6]. In this case, the terrain classes are modelled probabilistically using Dirichlet distributions (analogous to [3], [5]) and trained networks are then used to predict the likelihood of slipping using data collected offline.

C. Gap in the Literature

The methods discussed in the prior two subsections have two notable shortcomings. The first is that, while existing property estimation algorithms condition physical properties on semantic classifications, this conditioning of properties is immutable, meaning the property values of semantic classes cannot be updated during operation [3], [5], [6]. This is undesirable especially when such physical properties have intra-class variation or there are errors in these initial estimates. For example, the coefficient of friction of ice depends on the conditions of ice formation, outdoor temperature, and the presence of snow on the surface of the ice. Assuming that the coefficient of all ice is identical may result in unsafe or overly conservative behavior.

The second short-coming of these methods is the one-way conditioning of physical properties on semantic classes. This can be unduly limiting because visual information alone may not be sufficiently informative. For example, in Fig. 1 it is challenging for the semantic segmentation network to differentiate between snow and ice. We note that in this work we make the assumption that tactile information may be used to disambiguate visual uncertainty. Using a tactile sensor in this instance to estimate the friction coefficient can inform and improve the accuracy of the semantic classification. Using physical properties to inform semantic classification can enable the application of multiple sensing modalities while providing a means of verification for semantic predictions output by trained networks. This improves the accuracy of semantic predictions without requiring new training data and network retraining.

To address these short-comings, this paper proposes a method to *jointly* estimate semantic classifications and physical properties online using Bayesian inference. Semantic classifications from vision enable the prediction of physical properties, and tactile measurements enable the online update of physical property estimates. This in turn enables the semantic classification likelihoods to be updated.

III. Preliminaries

This section introduces notation, moments, and conjugate pairs. Vectors, written as columns, are typeset in bold and lowercase, while sets and matrices are typeset in uppercase. The element i of a vector $\boldsymbol{x}$ is denoted as x_i. An n-dimensional closed interval is denoted by $[a, b]^n$. Let p denote a normalized probability mass or density function. Following the convention of [33], we distinguish between different probability mass or density functions by the variable used as an input argument into each p. To avoid confusion, the term *multi-modal* refers to multiple sensing modalities, while *mixture* refers to a distribution with more than one mode. Likewise, we use *vision* to refer to RGB-D data.

A. Conjugate Pairs

In probability, conjugate distributions are a pair of distributions such that, given a prior and a likelihood, the posterior computed using Bayes' theorem belongs to the same family

of distributions as the prior [34]. This enables tractable computation of closed-form solutions for the posterior [35]. We introduce two such conjugate pairs here: the Dirichlet and Categorical distributions, and the Dirichlet Normal-Inverse-Gamma product and Gaussian mixture distributions.

1) Dirichlet Conjugate Prior: Let $z \in \{1, \dots, k\}$ denote a discrete random variable. The probability mass function of the Categorical distribution represents the probability that a sample z belongs to class i:

$$p(z = i|\boldsymbol{\theta}) = Cat(z = i|\boldsymbol{\theta}) = \theta_i, \tag{1}$$

where $i \in \{1, 2, \dots, k\}$, $\boldsymbol{\theta} \in [0, 1]^k$, and $\sum_i \theta_i = 1$.

The Dirichlet distribution is a continuous k-variate probability distribution that is parameterized by a vector $\boldsymbol{\alpha} \in \mathbb{R}^k_{\geq 0}$ with probability density:

$$p(\boldsymbol{\theta}|\boldsymbol{\alpha}) = Dir(\boldsymbol{\theta}|\boldsymbol{\alpha}) \tag{2}$$

$$= \frac{\Gamma(\sum_{j=1}^k \alpha_j)}{\sum_{j=1}^k \Gamma(\alpha_j)} \prod_{j=1}^k \theta_j^{\alpha_j - 1}, \tag{3}$$

where

$$\Gamma(\alpha_j) = \int_0^\infty x^{\alpha_j - 1} \exp(-x) dx. \tag{4}$$

The Dirichlet distribution is a conjugate prior to the Categorical distribution as is formalized in the following theorem whose proof is provided in Appendix A:

Theorem 1. *Let $\mathcal{Z} = \{z_1, \dots, z_n\}$ be a set of measurements drawn from a Categorical distribution, $p(z_j = i|\boldsymbol{\theta})$, and let the prior for $\boldsymbol{\theta}$ be a Dirichlet distribution, $p(\boldsymbol{\theta}|\boldsymbol{\alpha})$. The posterior computed using Bayes' theorem, $p(\boldsymbol{\theta}|\mathcal{Z}, \boldsymbol{\alpha})$, is also a Dirichlet distribution such that:*

$$p(\boldsymbol{\theta}|\mathcal{Z}, \boldsymbol{\alpha}) = Dir(\boldsymbol{\theta}|\tilde{\boldsymbol{\alpha}}), \tag{5}$$

where

$$\tilde{\alpha}_j = \alpha_j + \sum_{z_i \in \mathcal{Z}} 1\{z_i = j\}, \tag{6}$$

and $1\{z_i = j\}$ is equal to 1 when the expected class of measurement z_i is class j and is zero otherwise.

Theorem 1 enables the computation of the predictive posterior. In particular, one of the goals of this paper is to predict the probability that a new measurement belongs to class i given prior measurements $\mathcal{Z}$:

$$p(z = i|\mathcal{Z}, \boldsymbol{\alpha}) = p(z = i|\tilde{\boldsymbol{\alpha}}) \tag{7}$$

$$= \int_{\boldsymbol{\theta}} p(z = i|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\tilde{\boldsymbol{\alpha}}) \mathrm{d}\theta, \tag{8}$$

where $p(z = i|\boldsymbol{\theta})$ represents the Categorical likelihood and $p(\boldsymbol{\theta}|\tilde{\boldsymbol{\alpha}})$ the Dirichlet posterior probability. Computing this integral exactly is challenging. Fortunately, the theory of conjugate priors enables a closed-form solution [36, (3)]:

$$p(z = i|\tilde{\boldsymbol{\alpha}}) = \frac{\tilde{\alpha}_i}{\sum_{j=1}^k \tilde{\alpha}_j}. \tag{9}$$

2) Dirichlet Normal-Inverse-Gamma Conjugate Prior: Let $\psi \in \mathbb{R}$ denote a continuous random variable. A Gaussian mixture is a continuous probability density function

$$p(\psi|\Theta) = \sum_{i=1}^k w_i \mathcal{N}(\psi|\mu_i, \sigma_i^2), \tag{10}$$

where $\Theta = \{w_i, \mu_i, \sigma_i^2\}_{i=1}^k$ and $\sum_{i=1}^k w_i = 1$. Next, we present a candidate for the conjugate prior of the Gaussian mixture model by conditioning the parameters, Θ, on a set of hyperparameters, Ψ. Previously, the Dirichlet distribution was used to parameterize these variables to form a conjugate pair. We perform a similar trick here using the Dirichlet Normal-Inverse-Gamma product distribution as the prior on Θ.

The Dirichlet Normal-Inverse-Gamma product distribution is defined as

$$p(\Theta|\Psi) = Dir(\boldsymbol{w}|\boldsymbol{a}) \prod_{i=1}^k \mathcal{N}\Gamma^{-1}(\mu_i, \sigma_i^2|\tau_i, \kappa_i, \beta_i, \gamma_i), \tag{11}$$

where $\Psi = \{a_i, \tau_i, \kappa_i, \beta_i, \gamma_i\}_{i=1}^k$, $\kappa_i, \beta_i, \gamma_i > 0$, and the Normal-Inverse-Gamma distribution is defined as

$$\mathcal{N}\Gamma^{-1}(\mu_i, \sigma_i^2|\tau_i, \kappa_i, \beta_i, \gamma_i) = \frac{\sqrt{\kappa_i}}{\sqrt{2\pi\sigma_i^2}} \frac{\gamma_i^{\beta_i}}{\Gamma(\beta_i)} \frac{1}{\sigma_i^2}^{\beta_i + 1} \cdot \tag{12}$$

$$\cdot \exp(-\frac{2\gamma_i + \kappa_i(x - \tau_i)^2}{2\sigma_i^2}).$$

When a measurement ψ is taken, we assume it is drawn from the Gaussian mixture measurement likelihood. The posterior of Θ is then computed using the following theorem from [37, (6)]:

Theorem 2. *Let ψ be a measurement drawn from a Gaussian mixture $p(\psi|\Theta)$. Let the prior for Θ be a Dirichlet Normal-Inverse-Gamma product distribution. Then the posterior computed using Bayes' theorem, $p(\boldsymbol{\Theta}|\psi, \Psi)$, is:*

$$p(\Theta|\psi, \Psi) = \frac{1}{M} \sum_{j=1}^k c_j Dir(\boldsymbol{w}|\tilde{\boldsymbol{a}}_j) \cdot$$

$$\cdot \mathcal{N}\Gamma^{-1}(\mu_j, \sigma_j^2|\tilde{\tau}_j, \tilde{\kappa}_j, \tilde{\beta}_j, \tilde{\gamma}_j) \cdot$$

$$\cdot \prod_{i \neq j}^k \mathcal{N}\Gamma^{-1}(\mu_i, \sigma_i^2|\tau_i, \kappa_i, \beta_i, \gamma_i), \tag{13}$$

where

$$\tilde{a_j} = a_j + 1, \tag{14}$$

$$\tilde{\tau_j} = \frac{\kappa_j \tau_j + \psi}{\kappa_j + 1}, \tag{15}$$

$$\tilde{\kappa_j} = \kappa_j + 1, \tag{16}$$

$$\tilde{\beta}_j = \beta_j + \frac{1}{2}, \tag{17}$$

$$\tilde{\gamma_j} = \gamma_j + \kappa_j \frac{(\psi - \tau_j)^2}{2(1 + \kappa_j)}, \tag{18}$$

$$c_j = \sqrt{\frac{\kappa_j}{\tilde{\kappa_j}}} \frac{\Gamma(\tilde{\beta}_j)}{\Gamma(\beta_j)} \frac{\gamma_j^{(\beta_j)}}{\tilde{\gamma}_j^{(\tilde{\beta}_j)}}, \tag{19}$$

Algorithm 1: Method of Moments

Requires : Dataset $\mathcal{D} = \{\psi_l\}_{l=1}^N$, Prior $p(\Theta|\Psi)$

for $\psi \in \mathcal{D}$ **do**
- Compute $p(\Theta|\psi, \Psi)$ // (13)
- $\mathbb{E}[g_j(\boldsymbol{\theta})] \leftarrow$ moments$(p(\Theta|\psi, \Psi))$ // (20)
- $\hat{\Psi} \leftarrow$ matchMoments() // (22)–(26)
- Compute $p(\Theta|\hat{\Psi})$

$\tilde{\boldsymbol{a}}_j = \{a_1, \dots a_{j-1}, \tilde{a}_j, a_{j+1}, \dots, a_k\}$ *and* M *is the normalizing factor.*

Using Theorem 2, note the posterior, (13), is not in the same family of distributions as the prior, (11). More troublingly, the number of terms in (13) grows exponentially with the number of measurements. Rather than apply Theorem 2, we implement the method of moments described in the next section to approximate (13) as a Dirichlet Normal-Inverse-Gamma product distribution.

B. Method of Moments

Let $\boldsymbol{\lambda} = \{\lambda_1, \dots, \lambda_n\}$ be an n-dimensional continuous random variable with probability density function $p(\boldsymbol{\lambda}|\xi)$ conditioned on parameters ξ. The i-th order moment is defined as

$$M_{g_i(\boldsymbol{\lambda})}(p) = \mathbb{E}[g_i(\boldsymbol{\lambda})] \tag{20}$$

where g_i is a monomial of $\boldsymbol{\lambda}$ of degree i. For some distributions, there exists a finite set of such moments, termed the *sufficient moments* [37] that fully define the set of parameters ξ. This means that one can construct the probability density function for this distribution by just using the sufficient moments. We denote the set of sufficient moments of a distribution p as $\mathbb{S}_p$.

The method of moments is a technique that fits a distribution to a set of sampled data by matching the sufficient moments of the distribution with the empirical moments computed from data. This approach may also be used to approximate a distribution by another distribution in a different family of distributions by matching sufficient moments of the first distribution with those of the second distribution. For our purposes, we use the method of moments to approximate the posterior computed using Theorem 2 with a probability density from the family of Dirichlet Normal-Inverse-Gamma product distributions. This may be thought of as projecting the first distribution onto the family of the second distribution [37]. This process is summarized in Algorithm 1 for projecting (13) onto the family of distributions of (11).

C. Revisiting the Dirichlet Normal-Inverse-Gamma Product

By implementing Algorithm 1, we project (13) onto the family of Dirichlet Normal-Inverse-Gamma product distributions resulting in an approximate posterior. To compute this projection using the method of moments, we first need the sufficient moments of the Dirichlet Normal-Inverse-Gamma product distribution. From [37, §4.1], these sufficient moments are $\mathbb{S}_p = \{\mu_i, \sigma_i^2, \sigma_i^4, \mu_i^2\sigma_i^2, w_i, w_i^2\}_{i=1}^k$. This process is summarized in the following theorem whose proof is given in Appendix B:

Theorem 3. *Let ψ be a measurement drawn from a Gaussian mixture $p(\psi|\Theta)$. Let the prior for Θ be a Dirichlet Normal-Inverse-Gamma product distribution. Then the posterior computed using Bayes' theorem, $p(\boldsymbol{\theta}|\psi, \Psi)$, and projected onto the family of Dirichlet Normal-Inverse-Gamma product distributions via Algorithm 1 yields:*

$$p(\Theta|\hat{\Psi}) = Dir(\boldsymbol{w}|\hat{\boldsymbol{a}}) \prod_{i=1}^{k} \mathcal{N}\Gamma^{-1}(\mu_i, \sigma_i^2|\hat{\tau}_i, \hat{\kappa}_i, \hat{\beta}_i, \hat{\gamma}_i). \tag{21}$$

The parameters $\hat{\Psi}$ are computed using the sufficient moments of (13)*:*

$$\hat{\tau}_i = \mathbb{E}[\mu_i], \tag{22}$$

$$\hat{\kappa}_i = \frac{1}{\mathbb{E}[\mu_i^2\sigma_i^2] - \mathbb{E}[\mu_i]^2\mathbb{E}[\sigma_i^2]}, \tag{23}$$

$$\hat{\beta}_i = \frac{\mathbb{E}[\sigma_i^2]^2}{\mathbb{E}[\sigma_i^4] - \mathbb{E}[\sigma_i^2]^2}, \tag{24}$$

$$\hat{\gamma}_i = \frac{\mathbb{E}[\sigma_i^2]}{\mathbb{E}[\sigma_i^4] - \mathbb{E}[\sigma_i^2]^2}, \tag{25}$$

$$\hat{a}_i = \mathbb{E}[w_i]\frac{\mathbb{E}[w_i] - \mathbb{E}[w_i^2]}{\mathbb{E}[w_i^2] - \mathbb{E}[w_i]^2}. \tag{26}$$

Theorem 3 is applied sequentially when multiple measurements are given. We summarize this procedure in Algorithm 1. The posterior predictive distribution is then given as $p(\psi|\hat{\Theta})$, which is a Gaussian mixture, where

$$\hat{\Theta} = \mathbb{E}[p(\Theta|\hat{\Psi})]. \tag{27}$$

Algorithm 2: Semantic and Property Prediction

$\mathcal{M} \leftarrow$ initMap()
$\Theta \leftarrow$ initUncertainty()
while *robot is running* **do**
- $s \leftarrow$ getSemanticPointCloud()
- $\mathcal{M} \leftarrow$ updateMap(s) `// Sec.V`
- $\Theta \leftarrow$ updateSemanticUncertainty(s) `// Sec.V`
- **if** *user requests property measurement taken* **then**
  - $\psi \leftarrow$ getMeasurement()
  - $\mathcal{V} \leftarrow$ getMeasurementLocation() `// Sec.IX`
  - $\Theta \leftarrow$ measurementUpdate$(\psi, \mathcal{V})$ `// Alg.1`

IV. Algorithm Overview

The proposed algorithm takes as input a stream of RGB-D images and physical property measurements from a tactile sensor and outputs semantic classifications and physical property estimates which are modelled probabilistically. We provide an overview of the proposed method presented in Algorithm 2. We focus on tactile and visual modalities, however our method

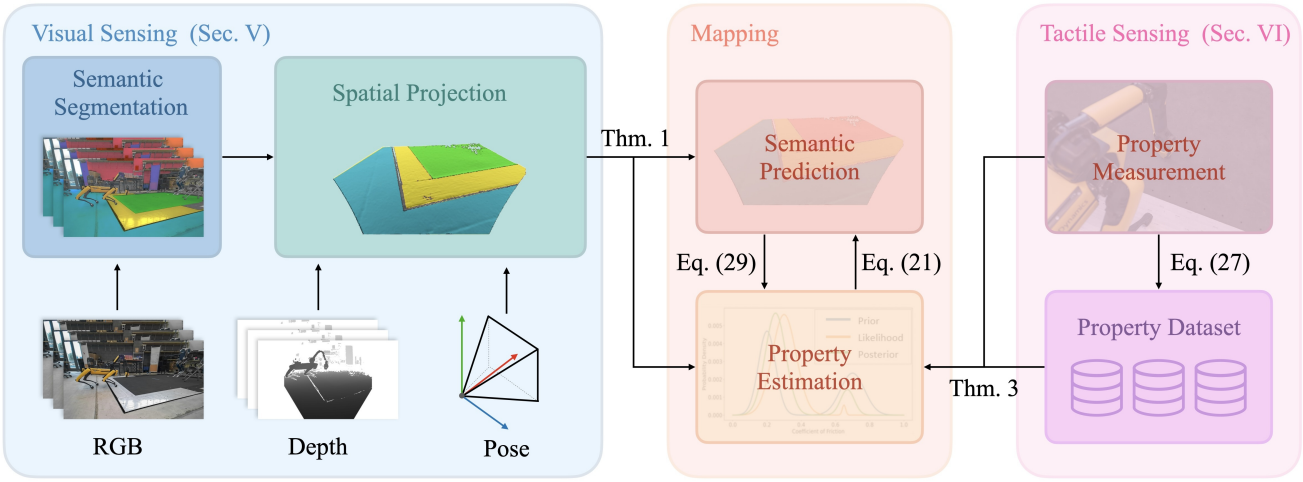

Fig. 2: A flow diagram illustrating Algorithm 2. A semantic classification algorithm predicts pixel-wise classes from RGB images that are then projected into a common mapping frame using the aligned depth image, camera intrinsics, and estimated camera pose. This semantic point cloud is used to build a metric-semantic map. When a property measurement is taken, Algorithm 1 is used to update the semantic and property estimates.

may be adapted to different sensing modalities in a straightforward manner.

A metric-semantic map and Dirichlet parameters are first initialized, where $\boldsymbol{\alpha} = \mathbf{1}_{k\times 1}$ (Lines 1 - 2). We assume that, while the robot is running, it is collecting exteroceptive data in the form of RGB-D images or lidar data. Using an off-the-self semantic segmentation network of choice, this data is transformed into a semantic point cloud [38] (Line 4). The geometric information of this point cloud is used to update the metric map (Line 5) while the semantic classifications are used to compute the semantic class posterior (Line 6), as described in Section V.

Locations for property measurements are specified by the user. When a property measurement is taken using a tactile sensor, we first compute the region over which the measurement was taken within the map and use moment matching (Algorithm 1) to update the semantic classifications and property estimates within this region (Lines 8-10). We describe this step in more detail in Section VI. A flow diagram illustrating this procedure is given in Figure 2.

V. Vision-Based Estimation

This section describes how to implement Bayesian inference to recursively estimate semantic classifications using visual data. This corresponds to the *Visual Sensing* block in Figure 2. First, we discuss how to obtain a semantic segmentation point cloud and then discuss how to probabilistically fuse this data within a geometric representation.

Semantic segmentation assigns task-dependent class probability scores to each pixel in an image. It is commonplace to use neural networks to estimate pixel-wise class probability scores [39]. Let $I \in \mathbb{R}^{w\times h\times 3}$ denote an RGB image, where h, w are the height and width of the image in pixels. A trained semantic segmentation network takes an input image I and outputs the pixel-wise class probability scores $T \in \mathbb{R}^{w\times h\times k}$ for k semantic classes in the form of a k-dimensional Categorical distribution. By employing a one-hot encoding over these pixel-wise scores, $H = 1\{T\}$ such that $H \in \mathbb{R}^{w\times h}$, we may interpret the network as drawing pixel-wise class samples from pixel-wise Categorical distributions. We denote these as the *semantic class measurements*, the accuracy of which depends on the network used.

We use an aligned depth image, camera intrinsics, and camera pose estimate to project the semantic class measurements H into a fixed global coordinate frame [38, (10.38)] to obtain a point cloud representation of the semantically segmented image. Akin to [40] we implement a modified version of Kinect Fusion [41], a truncated signed distance field (TSDF), as the geometric representation onto which semantic class measurements are projected. Within each voxel we store unique Dirichlet parameters $\boldsymbol{\alpha}$. Given new semantic class measurements, we update these parameters in each voxel using (6). Thus, semantic classification estimates may be recursively obtained with uncertainty estimates via (5).

VI. Tactile-Based Estimation

This section describes how semantic class estimates may be used to predict physical properties and how we use measurements of these properties to update semantic class likelihoods. This corresponds to the *Tactile Sensing* block in Figure 2. Our goal is to compute a measurement likelihood for physical properties using semantic class likelihoods introduced in Section V.

Motivated by [5], [42], we construct a conditional probability distribution $p(\psi \mid \mathcal{Z})$ for a property ψ with semantic data $\mathcal{Z}$. By applying the law of total probability and including the $\boldsymbol{\alpha}$ parameters introduced in Section III-A1, we derive the

Semantic Class	Gaussian Parameters	
	μ	σ
Concrete	0.543	0.065
Grass	0.577	0.077
Rock	0.478	0.133
Wood	0.372	0.055
Rubber	0.616	0.048
Plastic	0.311	0.045
Snow	0.390	0.071
Ice	0.192	0.046

TABLE I: Gaussian parameters for the coefficient of friction of semantic classes. These values are computed using the material friction dataset provided by [5] and assume rubber as the contact material.

following expression:

$$p(\boldsymbol{\psi} \mid \mathcal{Z}, \boldsymbol{\alpha}) = \sum_{i=1}^{k} p(\boldsymbol{\psi} \mid z = i) p(z = i \mid \mathcal{Z}, \boldsymbol{\alpha}). \tag{28}$$

There are two components to this equation. The first, $p(z = i \mid \mathcal{Z}, \boldsymbol{\alpha})$, is the posterior predictive likelihood of a region of the map belonging to semantic class i given prior semantic class measurements within that region. This posterior predictive likelihood was derived in Section III-A and is given by (9).

Physical properties are not constant across a semantic class and thus should be estimated via a distribution of possible values, not by a single value. The second component, $p(\boldsymbol{\psi} \mid z = i)$, denotes this likelihood distribution and is approximated as a Gaussian distribution [5]. Substituting (9) and the formula for the class-wise Gaussian model into (28) gives a closed-form estimate for the physical properties given semantic class measurements:

$$p(\boldsymbol{\psi} \mid \mathcal{Z}, \boldsymbol{\alpha}) = \sum_{i=1}^{k} \frac{\alpha_i}{\sum_{j=1}^{k} \alpha_j} \mathcal{N}(\mu_i, \sigma_i^2). \tag{29}$$

Note that this model is equivalent to the Gaussian mixture model presented in (10) where the weights are computed using the predictive posterior of the semantic classifications. We use this as the measurement likelihood for the properties of interest.

Given that the Gaussian mixture model is used as the measurement likelihood, we use the approximate conjugate prior of this distribution, the Dirichlet Normal-Inverse-Gamma product, to represent uncertainty in the semantic classifications, w_i, and class-wise property parameters, μ_i and σ_i^2. Note that the Gaussian mixture weights, w_i, are computed using the Dirichlet parameters $\boldsymbol{\alpha}$. After a property measurement is taken, Algorithm 1 is implemented to update the belief over these parameters via (21). This algorithm is applied sequentially if multiple measurements are taken.

Using Algorithm 1, the semantic class likelihood, w_i, as well as the property parameters themselves, μ_i and σ_i^2, are updated. Furthermore, the Dirichlet Normal-Inverse-Gamma product posterior provides a means of uncertainty quantification in the posterior estimates of the semantic class likelihoods and property values, unlike learning-based approaches.

VII. Implementation

Algorithms 1 and 2 are implemented in C++ and include a ROS interface which will be released upon final submission of the paper. Experiments are conducted on a laptop with a 5.4GHz i9-13900HX processor, 32GB of RAM, and an Nvidia RTX 4060 GPU.

We use a custom implementation of the SegFormer network [43] trained on the Dense Material Segmentation Dataset [12]. The output of the network is then post-processed with a segment-based voting scheme using FastSAM [30] in a similar manner to [44]. When a property measurement is taken, Algorithm 1 updates a region of the geometric representation segmented using FastSAM. This incentivizes regions with spatial proximity and visual similarity to be updated using a single measurement rather than requiring one measurement for each voxel, making the algorithm more efficient.

Table I is used to initialize the Gaussian parameters for the coefficient of friction for the relevant semantic classes. The values in Table I were computed using the material friction dataset provided by [5]. If a friction measurement is taken of a material class not present in Table I, Algorithm 1 predicts the class from Table I with the most similar friction values.

For hardware demonstrations, static friction measurements are taken using a force-torque sensor which measures contact forces during the motion of an end-effector against a surface. These forces are converted to friction measurements using the equation relating normal and tangential forces:

$$\psi = \frac{F_t}{F_n}, \tag{30}$$

where F_t is the force tangential to the contact normal, F_n is the force parallel to the contact normal, and ψ is the coefficient of friction. A low-pass filter is used to filter out measurement noise from the force-torque sensor.

VIII. Simulation Results

We validate our approach in simulation and demonstrate property measurements improve semantic predictions. The simulation experiments also validate the approximation accuracy of the moment matching method in Algorithm 1 in approximating the posterior via Theorem 3.

A. Dense Material Segmentation Dataset Evaluation

We use 1800 images and ground-truth semantic labels taken from the testing set of the Dense Material Segmentation Dataset [12] and the pre-trained semantic segmentation neural network described in Section VII is used to predict semantic labels. Locations of incorrect semantic classification are determined by comparing to the ground-truth semantic classifications provided by the dataset. Misclassified pixels in each image are then randomly selected and a friction measurement is drawn from a Gaussian distribution whose parameters are specified in Table I using the corresponding ground-truth semantic label. The parameters and classes in Table I are identical to those in [5] which were validated through thorough real-world experimentation. In particular, [5]

(a) Scene (b) Ground Truth (c) Only Vision (d) Ours

Fig. 3: Results for a single simulated experiment. The image (a) and ground truth semantic labels (b) are from the Dense Material Segmentation Dataset [12]. The semantic segmentation predictions (c) do not classify parts of the desk as wood. A friction measurement is simulated using the ground-truth semantic label and Table I and is sampled from the pixel highlighted by the red cross in the RGB image. Algorithm 1 then computes the correct posterior semantic label (d).

found that a Gaussian distribution using the parameters in Table I modeled real-world measurements well. Note that we use a similar prior model during our real-world experiments as are described in Section IX.

The semantic prediction is used to initialize the Dirichlet parameters such that $\boldsymbol{\alpha} = 1_{k\times 1} + 1\{z = j\}$. The Gaussian parameters are initialized via Table I and the remaining Dirichlet Normal-Inverse-Gamma parameters are initialized as $\boldsymbol{a} = \boldsymbol{\alpha}$, and $\tau_i = \mu_i$, $\kappa_i = 1$, $\beta_i = \sigma_i^2/\tau_i$, $\gamma_i = \sqrt{\beta_i}/C$, and $C = 40$ for each semantic class.

After a simulated friction measurement is randomly sampled, Algorithm 1 is invoked and the posterior predictive likelihood is computed. The Segment Anything Model [44] computes a mask over the input RGB image using the chosen pixel as the query point and the semantic predictions over this region are updated corresponding to the posterior predictive likelihood output by Algorithm 1. An example of this computed posterior is demonstrated in Figure 3.

B. Habitat Simulation Evaluation

We perform additional experiments in the Habitat simulator [45] which provides ground-truth RGB, depth, and semantic data for various indoor scenes. We demonstrate the performance of the Selmap [5] baseline using three pre-trained semantic segmentation models: the Dense Material Segmentation Dataset model [12], the BERT image transformer [46], and OneFormer [47]. We note that the Selmap baseline is itself not a semantic segmentation model, but rather a method for probabilistically fusing semantic segmentation images from these networks within a 3D map.

Incorrectly classified regions are chosen by comparing the predicted semantic classifications with the ground-truth semantic information from the Habitat simulator. We simulate friction measurements for incorrectly classified regions by randomly sampling from the Gaussian distribution whose parameters are specified in Table I using the corresponding ground-truth semantic label provided by the simulator. Algorithm 1 is then invoked to compute the posterior predictive likelihood of the semantic label.

We perform these simulation experiments for 5 distinct scenes in the Habitat simulator and use 10 RGB-D images per scene with known pose. Shown in Figure 4 are the results for Selmap alongside our proposed approach and ground-truth semantic labels for a sample scene from the Habitat simulator.

Table II contains the evaluation metrics, including peak signal-to-noise ratio (PSNR), structural similarity (SSIM), binary cross entropy, and mean pixel-wise accuracy, used to evaluate our proposed approach against the vision-only Selmap [5] baseline. These metrics are computed using both the Dense Material Segmentation Dataset results in Section VIII-A and the Habitat simulation results discussed above. We note that by leveraging tactile sensing, all evaluation metrics improve when compared to the vision-only baseline.

This experiment highlights that by conditioning semantic classifications on physical properties, one improve the accuracy of semantic classifications.

IX. Hardware Experiments

Next, we validate our method on several hardware demonstrations and compare against existing semantic mapping and property estimation approaches. In the first demonstration, we validate the results from simulation and show that our approach is capable of correcting erroneous semantic classifications by measuring the coefficient of friction. We then provide two case studies for the utility of estimating physical properties. For the first case study, our proposed approach is used within a risk-aware legged locomotion planner to update the locomotion gait of a quadruped when hazardous terrain is perceived. In the second case study, we demonstrate our approach is able to estimate affordance-based properties in a door-opening scenario. Both case studies are presented on the accompanying project webpage.

A. Semantic Prediction Using Property Measurements

The end-effector of a Kinova Gen 3 robotic arm and a Spot Arm are used to collect friction measurements using built-in wrist-mounted force-torque sensors. For this experiment, 5 frames from the RGB-D stream are used to initialize the semantic TSDF map. Each RGB-D image is semantically segmented using the trained network and projected into the global coordinate frame using the aligned depth image. The recursive vision-based semantic update described in Section V is applied for each semantic point cloud. This initializes the $\boldsymbol{\alpha}$ parameters in (29) used to compute the initial semantic classification weights for the measurement likelihood.

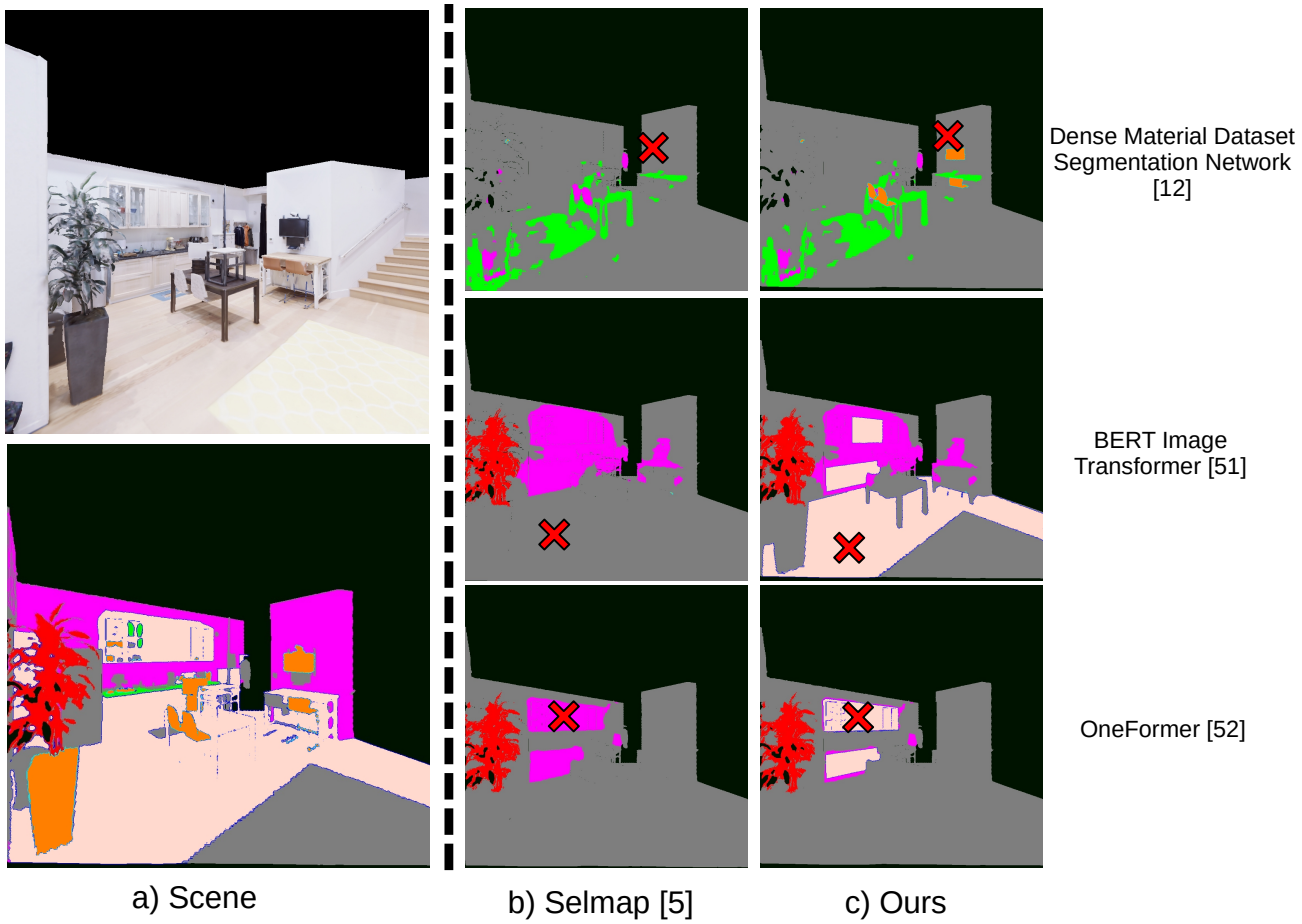

Fig. 4: Results for a single simulated experiment in the Habitat simulator [45]. a) The input RGB image with ground truth semantic labels is shown. On the right-hand side are the outputs of the Selmap [5] baseline and our proposed approach using three different semantic segmentation networks. b) The incorrect semantic labelling is reflected in the Selmap implementation as the semantic segmentation networks all predict incorrect classifications in various regions in the scene. By exploiting tactile measurements taken at the locations denoted by the red 'X', c) our proposed approach is able to correct these erroneous semantic predictions.

		Concrete	Grass	Rock	Wood	Rubber	Plastic	Snow	Ice	Total
Accuracy (↑)	Vision Only	0.70	0.74	0.70	0.76	0.81	0.82	0.92	0.82	0.78
	Ours	**0.83**	**0.77**	**0.79**	**0.82**	**0.83**	**0.85**	**0.99**	**0.84**	**0.83**
PSNR (↑)	Vision Only	8.48	11.91	11.90	8.97	12.01	11.26	16.66	11.12	9.96
	Ours	**11.59**	**12.07**	**13.43**	**10.44**	**12.07**	**12.07**	**21.76**	**11.44**	**11.20**
SSIM (↑)	Vision Only	0.70	0.73	0.69	0.74	**0.79**	0.80	0.92	0.81	0.76
	Ours	**0.80**	**0.75**	**0.77**	**0.79**	**0.79**	**0.82**	**0.98**	**0.82**	**0.80**
Binary Cross Entropy (↓)	Vision Only	2.72	2.67	2.73	2.58	2.86	2.53	1.60	2.69	2.58
	Ours	**2.29**	**2.61**	**2.45**	**2.40**	**2.83**	**2.46**	**1.39**	**2.63**	**2.43**

TABLE II: The accuracy, peak signal-to-noise ratio (PSNR), structural similarity (SSIM), and binary cross entropy metrics for the simulation experiment in Section VIII. The vision-only baseline uses the SegFormer+FastSAM network described in Section VII. Our proposed approach uses the same network for visual classification and uses simulated property measurements and Algorithm 1 to update semantic predictions. Property measurements are shown to improve semantic prediction for all material classes across all metrics.

The Gaussian parameters are initialized via Table I and the remaining Dirichlet Normal-Inverse-Gamma parameters are initialized as per Section VIII.

We compare our approach to the vision-only recursive semantic mapping approach of [5] which probabilistically filters semantic segmentation images within a metric map. The same SegFormer-FastSAM semantic segmentation network is used for both methods. As shown in Figure 5, the network incorrectly classifies various objects in the scenes. When only vision-based semantic classifications are considered [5], the erroneous predictions from the network are unable to be corrected (Figure 5c).

Using the proposed method, a user specifies the location for friction measurements to be taken (Figure 5a). Algorithm 1 then computes the posterior semantic classification weights using the friction measurements as input. This corrects the expected semantic classification, as shown in Figure 5d.

We ran the experiment on two indoor scenes with two friction measurements each as depicted in Figure 5. As shown in Figure 5, the semantic prediction accuracy increases using the proposed method and matches the ground-truth semantic labels. In contrast, [5] is unable to correct the erroneous predictions from the semantic segmentation network, resulting in incorrect semantic predictions. This demonstrates that the simulation results shown in Section VIII translate to hardware experiments.

B. Friction Estimation Using Semantics

We demonstrate the utility of our proposed property estimation method using a case study involving a challenging legged locomotion traversal task of crossing an icy surface. We implement a locomotion planner similar to [48] that switches between predetermined static and dynamic gaits depending on the expected value for the coefficient of friction of the terrain to be traversed. A static gait is used when the expected value for the coefficient of friction falls below a threshold of $\mathbb{E}[\psi] \leq 0.25$.

A stream of RGB-D images is used to build a semantic TSDF map of the robot's environment as outlined in Section V. The resultant semantic TSDF map predicts that the region the robot must traverse may have a low coefficient of friction and a measurement is therefore taken to better approximate the value.

The Dirichlet Normal-Inverse-Gamma prior parameters are initialized as described in Section VIII. We use the manipulator mounted to the legged robot to take friction measurements of the ground in front of the robot using a built-in wrist-mounted force-torque sensor via (30). Algorithm 1 then computes the posterior estimate for the coefficient of friction (Figure 1).

The experiment was conducted six times in the same location during the same afternoon, three where the robot does not measure the coefficient of friction and three where it does. The prior and posterior friction estimates for these six trials is shown in Figure 6. The Selmap baseline incorrectly predicts the semantic label of the terrain for each run and misclassifies the ice as either concrete or snow. Since a friction measurement is not taken, the friction threshold $\mathbb{E}[\psi] \leq 0.25$ is not met due to mislabeling and the robot attempts to cross the ice using a dynamic gait, resulting in it slipping and falling for all three of these trials. This comparison can be found at the project page website[2].

For the subsequent two trials, friction measurements of $\psi_1 = 0.139$ and $\psi_2 = 0.156$ are taken for each trial, respectively. After computing the posterior friction estimates via Algorithm 1, the friction threshold is met and the robot switches to a static, stable gait to cross the ice. This results in successful ice traversal for these two trials. In the following third trial, a friction measurement of $\psi_3 = 0.628$ is taken due to erroneous end-effector placement. The computed posterior is shown in Figure 6 and shows a coefficient of friction with a large expected value which exceeds the threshold. For this trial, the robot remains in a dynamic gait and slips on the ice during traversal. A video of a successful ice traversal trial is provided on the accompanying project webpage.

We further compare our property estimation approach with two state-of-the-art traversability estimation methods designed for legged locomotion. Both methods consider both geometric and semantic features and use a pre-trained neural network to estimate a traversability score [3], [49]. Both traversability estimation methods predict that the icy surface is traversable as shown in Figure 7; however, attempting to traverse this surface using a dynamic gait resulted in the robot slipping and falling.

Our proposed method is able to verify the property predictions estimated using visual information via tactile sensing and Bayesian inference whereas other state-of-the-art property estimation methods are unable to incorporate multiple sensing modalities. This task highlights the utility of estimating physical properties such as the coefficient of friction to accomplish challenging tasks such as legged locomotion traversal across an icy surface. State-of-the-art traversability estimation methods depend only on the terrain geometry and classification and thus do not provide the means by which a robot may adapt its traversal strategy to account for the properties (i.e. friction) which dictate traversability.

C. Encoding Affordance-Based Properties

Lastly, we present a scenario in which our approach is used to estimate affordance-based properties. An affordance is a property of an object that defines its possible used [50]. In this case, we focus on the affordance of a door handle, which affords either a push or pull action, as the semantic classification and examine the force required to open the door as the property of interest.

As shown in Figure 8, a mobile manipulator is used to measure the force required to open a door. Negative forces constitute pushing while positive forces constitute pulling. We equally weight each push/pull affordance and initialize a mean force magnitude of $20N$ and a variance of $10N$.

To measure the force required to open the door, the robot begins applying a pulling force $10N$ and ramps this up to $70N$ over a period of 3 seconds. At $57N$ the door begins to open and this value is used as the property measurement for Algorithm 1. The door opening force posterior is show in Figure 8 in the lower right-hand corner. After invoking Algorithm 1 the push/pull affordances are still almost equally weighed. Additionally, the variance for the pulling force mode has increased, mirroring the fact that a higher-than-expected force was required to open the door. This door opening experiment is demonstrated on the accompanying project webpage.

Next, we test the force required to open six different doors, three pull-to-open and three push-to-open, and use Algorithm

[2]https://roahmlab.github.io/multimodal_mapping/

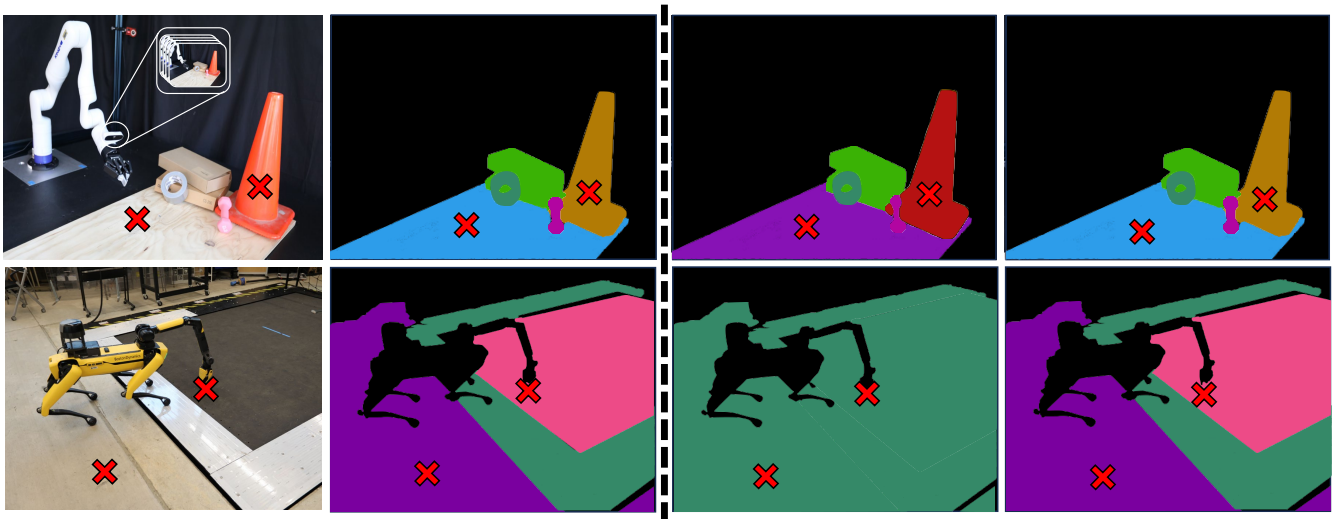

Fig. 5: A semantic segmentation task is shown and the proposed method is compared against the semantic mapping approach from [5] which is called Selmap. The expected semantic class is shown. a) The input image with measurement locations shown using an X and associated b) ground truth semantic labels are provided. Both Selmap and our approach use the same SegFormer + FastSAM pre-trained semantic segmentation network specified in Section VII. This network incorrectly predicts the semantic labels for several regions within the scene. c) This incorrect labelling is reflected in the Selmap implementation as the visual-based semantic mapping approach is unable to correct these erroneous predictions. By exploiting a tactile sensing modality, d) our approach is able to correct the erroneous semantic predictions and correctly predict the semantic labels of the objects within the scene.

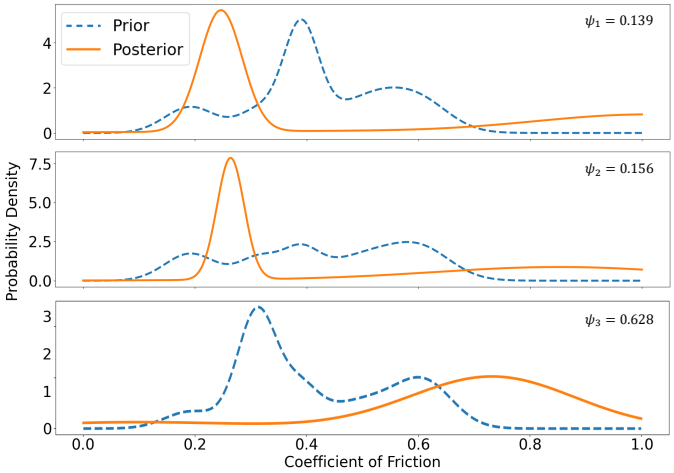

Fig. 6: The prior (dotted) and posterior (solid) friction estimates for the three ice traversal demonstration runs. When friction measurements are taken of the ice before traversal, Algorithm 1 predicts the posterior friction estimates. The friction measurements are $\psi_1 = 0.139$, $\psi_2 = 0.156$, and $\psi_3 = 0.628$ for trials 1, 2, and 3, respectively.

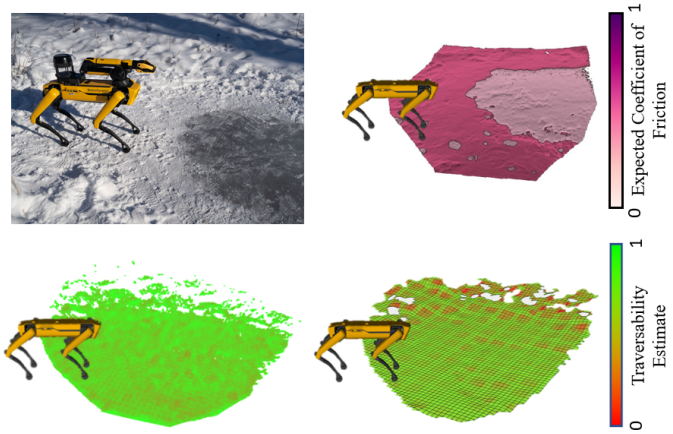

Fig. 7: The performance of the proposed approach after a friction measurement is taken in comparison to the traversability estimate computed by [49] (bottom left) and [3] (bottom right). The proposed method is able to correctly predict that the surface in front of the robot has a low coefficient of friction which informs the locomotion planner to switch to a stable gait. In contrast, the traversability estimates predict the ice is traversable even when the robot is using a dynamic gait, which results in the robot slipping and falling.

1 to compute the posterior door opening force estimate sequentially after each trail. Figure 9 shows the prior and posterior opening force estimate after each trial. After six trials the estimated door opening force posterior has maintained two modes corresponding to the push or pull force modalities, however the variance has increased, representing the varied door opening forces seen in the experiments compared to the initial estimate.

This case study demonstrates the utility of the proposed method for affordance-based property estimation and opens an avenue for future research given recent advances in affordance-based planning [51].

X. Discussion and Conclusion

We propose a method for jointly estimating semantic labels and physical properties. We model semantic classifications and physical properties probabilistically enabling closed-form Bayesian inference given visual semantic predictions and tactile property measurements. We demonstrate that by leveraging this multi-modal probabilistic approach we outperform the vision-only baselines across all evaluation metrics in simulation and provide hardware experiments demonstrating the

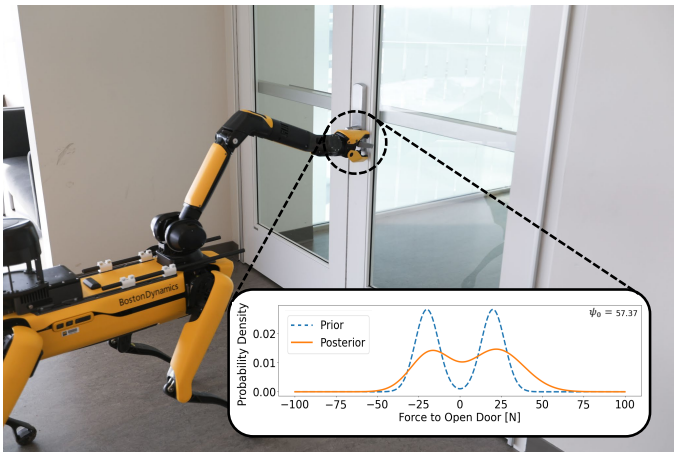

Fig. 8: The proposed method is capable of encoding affordance-based properties, in this case the force required to open a door. This door opening force is initialized as a multi-modal distribution with pushing and pulling being equally likely with a magnitude of 20N. The robot applies a pulling force of 57N on the door, resulting in it swinging open. The posterior estimate for the force required to open a door is computed and shown in the embedded plot.

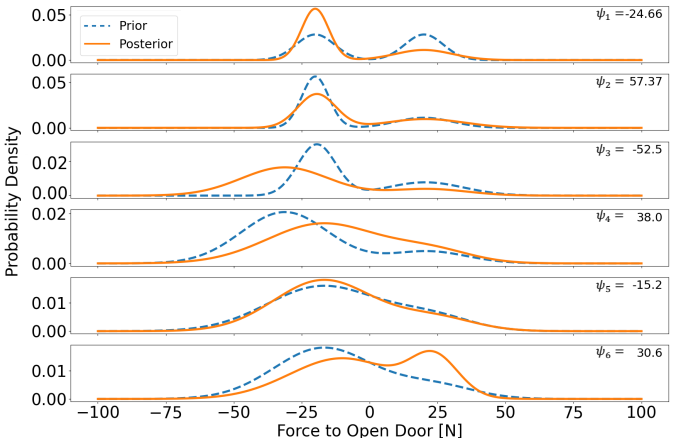

Fig. 9: The prior and posterior door opening force estimates running Algorithm 1 sequentially with six experiments, meaning the posterior for the previous experiment is the prior for the current experiment. The force required to open the door is indicated in the upper right-hand corner for each experiment. After measuring the force required to open 6 doors, the posterior estimate has a high variance with an expected value near $0N$, reflecting the equal likelihood that a door is either push or pull to open.

efficacy of the proposed approach. Additionally, we provide two case studies which motivate the utility of physical property estimation demonstrated in the proposed method.

One drawback of our approach is that the accuracy of our method is reliant on the accuracy of the semantic classification network. If a semantic classification network consistently predicts incorrect semantic labels over a long duration, the initial Dirichlet parameters will skew heavily towards the incorrect semantic label and be difficult to correct with property measurements. Additionally, like many voxel-based semantic mapping frameworks, our method is constrained by memory limitations which restrict the size of the map. We aim to address these concerns in future work.

Tactile and visual sensing modalities were used for the proposed method, however additional sensing modalities may also be used in place of or in addition to tactile sensing. This is an avenue for future work, along with incorporating tactile sensing noise. Motivated by recent work in legged locomotion [52], we aim to use the property estimates computed using our method as inputs to locomotion controllers. Such a supervisory signal would enable robots to change their gait before transitioning between surfaces and potentially enable safer, robust locomotion policies. A similar use for property uncertainty may be found in assessing locomotion risk, and a risk threshold may be used to determine when new tactile or visual measurements are required.

Lastly, our presented method provides a link between visual and tactile sensing modalities through physical property estimates and future work will aim to exploit this relationship for active perception. Current active perception methods are primarily concerned with geometric reconstruction, however as demonstrated in Section IX, physical properties play an important role in robotic task completion. As such, future work will explore how to leverage uncertainty in semantic and property estimates to find where in the environment a robot should explore, or where to take tactile measurements to decrease semantic uncertainty computed from visual data.

References

[1] Y. Guo, Y. Liu, T. Georgiou, and M. S. Lew, "A review of semantic segmentation using deep neural networks," *International journal of multimedia information retrieval*, vol. 7, pp. 87–93, 2018.

[2] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[3] L. Gan, Y. Kim, J. W. Grizzle, *et al.*, "Multitask learning for scalable and dense multilayer Bayesian map inference," *IEEE Transactions on Robotics*, vol. 39, no. 1, pp. 699–717, 2022.

[4] J. McCormac, A. Handa, A. Davison, and S. Leutenegger, "Semanticfusion: Dense 3d semantic mapping with convolutional neural networks," in *2017 IEEE International Conference on Robotics and automation (ICRA)*, IEEE, 2017, pp. 4628–4635.

[5] P. Ewen, A. Li, Y. Chen, S. Hong, and R. Vasudevan, "These maps are made for walking: Real-time terrain property estimation for mobile robots," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 7083–7090, 2022.

[6] X. Cai, S. Ancha, L. Sharma, *et al.*, "Evora: Deep evidential traversability learning for risk-aware off-road autonomy," *arXiv preprint arXiv:2311.06234*, 2023.

[7] S. Hao, Y. Zhou, and Y. Guo, "A brief survey on semantic segmentation with deep learning," *Neurocomputing*, vol. 406, pp. 302–321, 2020.

[8] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 7262–7272.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, Ieee, 2009, pp. 248–255.

[10] T.-Y. Lin, M. Maire, S. Belongie, *et al.*, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, Springer, 2014, pp. 740–755.

[11] B. Kuipers and Y.-T. Byun, "A robot exploration and mapping strategy based on a semantic hierarchy of spatial representations," *Robotics and autonomous systems*, vol. 8, no. 1-2, pp. 47–63, 1991.

[12] P. Upchurch* and R. Niu*, "A dense material segmentation dataset for indoor and outdoor scene parsing," in *ECCV*, 2022.

[13] I. Kostavelis and A. Gasteratos, "Semantic mapping for mobile robotics tasks: A survey," *Robotics and Autonomous Systems*, vol. 66, pp. 86–103, 2015.

[14] N. Sünderhauf, T. T. Pham, Y. Latif, M. Milford, and I. Reid, "Meaningful maps with object-oriented semantic mapping," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2017, pp. 5079–5085.

[15] J. Wilson, Y. Fu, J. Friesen, *et al.*, "ConvBKI: Real-time probabilistic semantic mapping network with quantifiable uncertainty," *arXiv preprint arXiv:2310.16020*, 2023.

[16] A. Hornung, K. M. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard, "OctoMap: An efficient probabilistic 3D mapping framework based on octrees," *Autonomous Robots*, 2013, Software available at https://octomap.github.io.

[17] C. Couprie, C. Farabet, L. Najman, and Y. LeCun, "Indoor semantic segmentation using depth information," *arXiv preprint arXiv:1301.3572*, 2013.

[18] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. L. Yuille, "Towards unified depth and semantic prediction from a single image," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2800–2809.

[19] Y. Sun, W. Zuo, P. Yun, H. Wang, and M. Liu, "FuseSeg: Semantic segmentation of urban scenes based on RGB and thermal data fusion," *IEEE Transactions on Automation Science and Engineering*, vol. 18, no. 3, pp. 1000–1011, 2020.

[20] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, Springer, 2012, pp. 746–760.

[21] D.-K. Kim, D. Maturana, M. Uenoyama, and S. Scherer, "Season-invariant semantic segmentation with a deep multimodal network," in *Field and Service Robotics: Results of the 11th International Conference*, Springer, 2018.

[22] A. Valada, R. Mohan, and W. Burgard, "Self-supervised model adaptation for multimodal semantic segmentation," *International Journal of Computer Vision*, vol. 128, no. 5, 2020.

[23] M. Brandao, K. Hashimoto, and A. Takanishi, "Friction from vision: A study of algorithmic and human performance with consequences for robot perception and teleoperation," in *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*, IEEE, 2016, pp. 428–435.

[24] D. Noh, H. Nam, M. S. Ahn, *et al.*, "Surface material dataset for robotics applications (SMDRA): A dataset with friction coefficient and RGB-D for surface segmentation," in *2020 25th International Conference on Pattern Recognition (ICPR)*, IEEE, 2021, pp. 6275–6281.

[25] S. Wang, *Road terrain classification technology for autonomous vehicle*. Springer, 2019.

[26] P. Papadakis, "Terrain traversability analysis methods for unmanned ground vehicles: A survey," *Engineering Applications of Artificial Intelligence*, vol. 26, no. 4, pp. 1373–1385, 2013.

[27] M. Shneier, W. Shackleford, T. Hong, and T. Chang, "Performance evaluation of a terrain traversability learning algorithm in the DARPA LAGR program," in *Performance Metrics for Intelligent Systems Workshop, Gaithersburg, MD, USA*, 2006, pp. 103–110.

[28] J. Frey, D. Hoeller, S. Khattak, and M. Hutter, "Locomotion policy guided traversability learning using volumetric representations of complex environments," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2022, pp. 5722–5729.

[29] D. Kim, J. Sun, S. M. Oh, J. M. Rehg, and A. F. Bobick, "Traversability classification using unsupervised on-line visual learning for outdoor robot navigation," in *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.*, IEEE, 2006, pp. 518–525.

[30] J. Frey, M. Mattamala, N. Chebrolu, C. Cadena, M. Fallon, and M. Hutter, "Fast traversability estimation for wild visual navigation," *arXiv preprint arXiv:2305.08510*, 2023.

[31] C. Sferrazza, Y. Seo, H. Liu, Y. Lee, and P. Abbeel, "The power of the senses: Generalizable manipulation from vision and touch through masked multimodal learning," *arXiv preprint arXiv:2311.00924*, 2023.

[32] G. B. Margolis, X. Fu, Y. Ji, and P. Agrawal, "Learning to see physical properties with active sensing motor policies," *arXiv preprint arXiv:2311.01405*, 2023.

[33] S. Thrun, "Probabilistic robotics," *Communications of the ACM*, vol. 45, no. 3, pp. 52–57, 2002.

[34] P. Diaconis and D. Ylvisaker, "Conjugate priors for exponential families," *The Annals of statistics*, pp. 269–281, 1979.

[35] B. D. Anderson and J. B. Moore, *Optimal filtering*. Courier Corporation, 2012.

[36] S. Tu, "The Dirichlet-Multinomial and Dirichlet-Categorical models for Bayesian inference," *Computer Science Division, UC Berkeley*, vol. 2, 2014.

[37] P. Jaini and P. Poupart, "Online and distributed learning of gaussian mixture models by bayesian moment matching," *arXiv preprint arXiv:1609.05881*, 2016.

[38] M. S. Nixon and A. S. Aguado, "Feature extraction & image processing for computer vision (third edition)," in *Feature Extraction & Image Processing for Computer Vision (Third Edition)*, M. S. Nixon and A. S. Aguado, Eds., Third Edition, Oxford: Academic Press, 2012, pp. 489–518.

[39] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez, and J. Garcia-Rodriguez, "A survey on deep learning techniques for image and video semantic segmentation," *Applied Soft Computing*, vol. 70, pp. 41–65, 2018.

[40] Y. Xiang and D. Fox, "DA-RNN: Semantic mapping with data associated recurrent neural networks," *arXiv preprint arXiv:1703.03098*, 2017.

[41] S. Izadi, D. Kim, O. Hilliges, *et al.*, "Kinectfusion: Real-time 3d reconstruction and interaction using a moving depth camera," in *Proceedings of the 24th annual ACM symposium on User interface software and technology*, 2011, pp. 559–568.

[42] T. Nguyen, F. Verdoja, F. Abu-Dakka, and V. Kyrki, "Probabilistic surface friction estimation based on visual and haptic measurements," *IEEE Robotics and Automation Letters*, vol. PP, pp. 1–1, Feb. 2021.

[43] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 077–12 090, 2021.

[44] J. Chen, Z. Yang, and L. Zhang, *Semantic segment anything*, https://github.com/fudan-zvg/Semantic-Segment-Anything, 2023.

[45] M. Savva, A. Kadian, O. Maksymets, *et al.*, "Habitat: A platform for embodied ai research," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9339–9347.

[46] H. Bao, L. Dong, and F. Wei, "Beit: BERT pre-training of image transformers," 2021.
[47] J. Jain, J. Li, M. Chiu, A. Hassani, N. Orlov, and H. Shi, "OneFormer: One Transformer to Rule Universal Image Segmentation," *arXiv*, 2022.
[48] P. Filitchkin and K. Byl, "Feature-based terrain classification for littledog," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 2012, pp. 1387–1392.
[49] R. Agishev, T. Petříček, and K. Zimmermann, "Trajectory optimization using learned robot-terrain interaction model in exploration of large subterranean environments," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3365–3371, 2022.
[50] J. J. Gibson, "The theory of affordances," *Hilldale, USA*, vol. 1, no. 2, 1977.
[51] S. Rezapour Lakani, A. J. Rodríguez-Sánchez, and J. Piater, "Towards affordance detection for robot manipulation using affordance for parts and parts for affordance," *Autonomous Robots*, vol. 43, pp. 1155–1172, 2019.
[52] T. Miki, J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, "Learning robust perceptive locomotion for quadrupedal robots in the wild," *Science Robotics*, vol. 7, no. 62, 2022.
[53] T. Minka, *Estimating a Dirichlet distribution*, 2000.
[54] J. M. Bernardo and A. F. Smith, *Bayesian theory*. John Wiley & Sons, 2009, vol. 405.

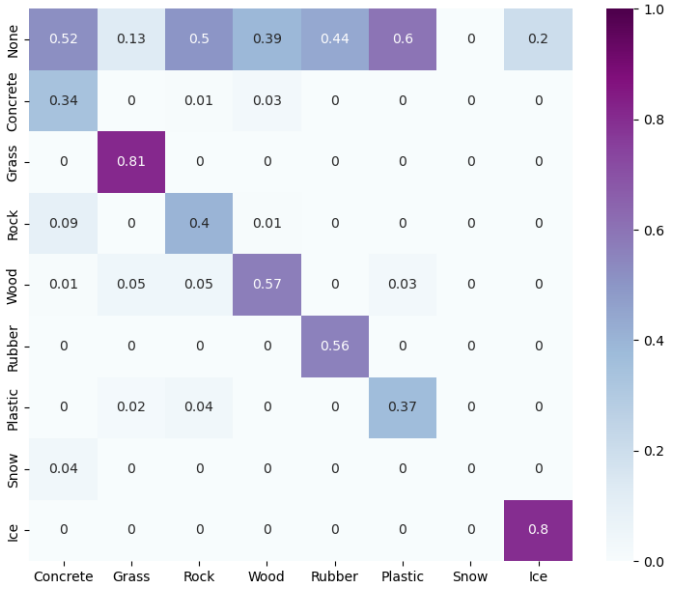

Fig. 10: Confusion matrix for the combined SegFormer+FastSAM semantic segmentation network computed using a randomly chosen subset of the test set of the Dense Material Segmentation Dataset [12]. There are no images of snow in this subset, as demonstrated by the confusion matrix.

Appendix A
Proof of Theorem 1

Let $p(\boldsymbol{\theta}|\boldsymbol{\alpha})$ be the prior belonging to the family of Dirichlet distributions and $p(z|\boldsymbol{\theta})$ the likelihood belonging to the family of Categorical distributions. Let $\mathcal{Z} = \{z_1, \ldots, z_n\}$ be a set of measurements drawn from the Categorical likelihood.

The Law of Total Probability is used to compute the posterior distribution:

$$p(\boldsymbol{\theta}|\mathcal{Z}, \boldsymbol{\alpha}) \propto p(\boldsymbol{\theta}, \mathcal{Z}, \boldsymbol{\alpha}) \tag{31}$$

$$= p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \prod_{z_i \in \mathcal{Z}} p(z_i|\boldsymbol{\theta}), \tag{32}$$

where the equality is due to conditional independence.

Then, substituting (1) and (2) into (32), we arrive at:

$$p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \prod_{z_i \in \mathcal{Z}} p(z_i|\boldsymbol{\theta}) \propto \prod_{j=1}^{k} \theta_j^{\alpha_j - 1} \prod_{z_i \in \mathcal{Z}} \prod_{j=1}^{k} \theta_j^{1\{z_i = j\}} \tag{33}$$

$$= \prod_{j=1}^{k} \theta_j^{\alpha_j - 1 + \sum_{z_i \in \mathcal{Z}} 1\{z_i = j\}}. \tag{34}$$

This is exactly the Dirichlet distribution with parameters

$$\tilde{\alpha}_j = \alpha_j + \sum_{z_i \in \mathcal{Z}} 1\{z_i = j\}. \tag{35}$$

Appendix B
Proof of Theorem 3

Let $p(\Theta|\Psi)$ be the prior belonging to the family of Dirichlet Normal-Inverse-Gamma distributions and $p(\psi|\Theta)$ the likelihood belonging to the family of Gaussian mixtures. Let ψ be a measurements drawn from the Gaussian mixture likelihood.

From Theorem 2, the posterior of Θ is given by:

$$p(\Theta|\psi, \Psi) = \frac{1}{M} \sum_{j=1}^{k} c_j Dir(\boldsymbol{w}|\tilde{\boldsymbol{a}}_j) \cdot \cdot \mathcal{N}\Gamma^{-1}(\mu_j, \sigma_j^2|\tilde{\tau}_j, \tilde{\kappa}_j, \tilde{\beta}_j, \tilde{\gamma}_j) \cdot \cdot \prod_{i \neq j}^{k} \mathcal{N}\Gamma^{-1}(\mu_i, \sigma_i^2|\tau_i, \kappa_i, \beta_i, \gamma_i), \tag{36}$$

with parameters computed as per (14)-(19).

It is evident that the number of terms in this posterior grows exponentially for each measurement. To overcome this challenge, we project (36) onto the family of Dirichlet Normal-Inverse-Gamma product distributions using the method of moments. We now prove the set of sufficient moments for the Dirichlet Normal-Inverse-Gamma product distribution is $\mathbb{S} = \{\mu_i, \sigma_i^2, \sigma_i^4, \mu_i^2\sigma_i^2, w_i, w_i^2\}_{i=1}^{k}$.

From (21), it is evident that $\boldsymbol{w}$ drawn from the Dirichlet distribution and $\Phi_i = \{\mu_i, \sigma_i\}$ drawn from the Normal-Inverse-Gamma distributions are independent random variables. Therefore, to find the sufficient moments of the Dirichlet Normal-Inverse-Gamma product distribution, we may use the sufficient moments for the Dirichlet and Normal-Inverse-Gamma distributions. The sufficient moments for the Dirichlet distribution are $\{\boldsymbol{w}, \boldsymbol{w}^2\}$ [53] and the sufficient moments of the i-th Normal-Inverse-Gamma distribution are $\{\mu_i, \sigma_i^2, \sigma_i^4, \mu_i^2\sigma_i^2\}$ [54, Chapter 4.5]. The concatenation of these sufficient moments then constitute the sufficient moments of (21).

The parameters of the Dirichlet Normal-Inverse-Gamma product distribution are then computed via (14)-(19) using the sufficient statistics as derived in [53] and [54], respectively.

A. Computing the Sufficient Statistics

The sufficient statistics $\mathbb{S} = \{\mu_i, \sigma_i^2, \sigma_i^4, \mu_i^2\sigma_i^2, w_i, w_i^2\}_{i=1}^k$ for (21) must be computed from (36). It is possible to use numerical integration to compute the sufficient moments via (20), however we demonstrate a more computationally efficient method here by exploiting the linearity of the expectation operator and the independence of variables within (36).

We derive the first sufficient moment $\mathbb{E}[\mu_i]$ and leave the remaining moment derivations to the reader. We start by noting the independence of $\boldsymbol{w}$ and each set of parameters $\Phi_i = \{\mu_i, \sigma_i^2\}$. Thus, to compute $\mathbb{E}[\mu_i]$, we begin by marginalizing out all independent variables from (36), resulting in:

$$\begin{aligned}\mathbb{E}[\mu_i] &= \int_{\mu_i} \frac{\mu_i}{M}\Big[c_i\mathcal{N}\Gamma^{-1}(\mu_i,\sigma_i^2|\tilde{\tau}_i,\tilde{\kappa}_i,\tilde{\beta}_i,\tilde{\gamma}_i)+ \\ &\quad + \sum_{j\neq i}^{k} c_j\mathcal{N}\Gamma^{-1}(\mu_i,\sigma_i^2|\tau_i,\kappa_i,\beta_i,\gamma_i)\Big]\mathrm{d}\mu_i \qquad (37)\\ &= \frac{1}{M}\Big[c_i\int_{\mu_i}\mu_i\mathcal{N}\Gamma^{-1}(\mu_i,\sigma_i^2|\tilde{\tau}_i,\tilde{\kappa}_i,\tilde{\beta}_i,\tilde{\gamma}_i)\mathrm{d}\mu_i+ \\ &\quad + \sum_{j\neq i}^{k} c_j\int_{\mu_i}\mu_i\mathcal{N}\Gamma^{-1}(\mu_i,\sigma_i^2|\tau_i,\kappa_i,\beta_i,\gamma_i)\mathrm{d}\mu_i\Big] \qquad (38)\\ &= \frac{1}{M}\Big[c_i\tilde{\tau}_i + \sum_{j\neq i}^{k} c_j\tau_i\Big] \qquad (39)\end{aligned}$$

where the first equality follows from the linearity of the expectation operator and the second equality follows from the the fact that for a Normal-Inverse-Gamma distribution, $\mathcal{N}\Gamma^{-1}(\mu,\sigma^2|\tau,\kappa,\beta,\gamma)$, the expected value for μ is simply $\mathbb{E}[\mu] = \tau$.

APPENDIX C
ADDITIONAL IMPLEMENTATION DETAILS

This section describes the calibration of the uncertainty of the semantic segmentation network and the class-wise Gaussian model used to probabilistically relate material class to friction.

The pre-trained SegFormer+FastSAM network discussed in Section VII is evaluated on the same test set from the Dense Material Segmentation Dataset [12] used in the simulation experiments in Section VIII-A. The confusion matrix for this evaluation is shown in Figure 10. From this confusion matrix we note the semantic segmentation network struggles to correctly predict plastic, concrete, and grass, however it is able to accurately predict the other material classes with high likelihood. Notably, the biggest source of error in the network is from incorrectly predicting the *None* class rather than classification as a different material.

Only a subset of the Dense Material Segmentation Dataset classes are considered in this work. The remaining classes are clustered into the *None* category, resulting in a class imbalance in the dataset. This is the most likely cause for the incorrect *None* class predictions.

With respect to the Gaussian property model calibration, we refer to [5, §VII] which describes how this class-wise Gaussian model was chosen and calibrated.