Multi-Modal Semantic Perception Using Bayesian Inference

Parker Ewen, Gitesh Gunjal, Hao Chen, Anran Li, Yuzhen Chen, and Ram Vasudevan

***Abstract*—Semantic classification has become a popular method for scene understanding in the robotics community. For many robotic tasks such as manipulation, legged locomotion, or high-speed autonomous driving, semantic classifications offer important information regarding the physical properties on which these tasks are dependent. Unfortunately, using visual information alone can lead to incorrect semantic classifications leading to incorrect property estimation. In this work we demonstrate that by leveraging multiple sensing modalities we are able to better predict semantic classifications and their physical properties. We accomplish this by showing that material properties are conditioned on their semantic class and, likewise, semantic classifications may also be conditioned on their material properties. We demonstrate our approach on several real-world scenes where friction measurements are used to correct erroneous semantic classifications.**

I. Introduction

Semantic classification has received enormous attention over the past decade as a means of interpreting an agent's environment. Vision-based semantic classification via neural networks have led to the substantial advancements in semantic perception, yet the accuracy of these methods is reliant on the statistics of the training dataset [1]. Even more challenging, tasks such as camouflage detection [2] or even material classification may be ill-posed when considering visual information alone.

This work introduces a semantic classification framework primary suited to embodied agents with multiple sensing modalities where vision is one of these modalities. Our key insight is that semantic classifications are conditioned on their physical properties and physical properties may likewise be conditioned on the semantic classifications. We took a first step towards leveraging this insight for recursive terrain property estimation in our previous paper [3] using only visual information. Our method works by projecting semantically classified images onto a static spatial representation and applying Bayesian inference to recursively filter the semantic labels within the representation. Here we extend this framework to enable multiple sensing modalities to be leveraged in conjunction with vision for semantic classification using Bayesian inference.

We demonstrate our approach for multi-modal semantic classification and property estimation on several indoor and outdoor real-world scenes. We show our method is able to leverage the relationship between semantic classes and semantic properties to first make predictions using vision and update these predictions using additional sensing modalities. For these experiments we primarily focus on friction measurements, although additional physical properties (e.g., soft contact parameters, stability, etc.) are also applicable.

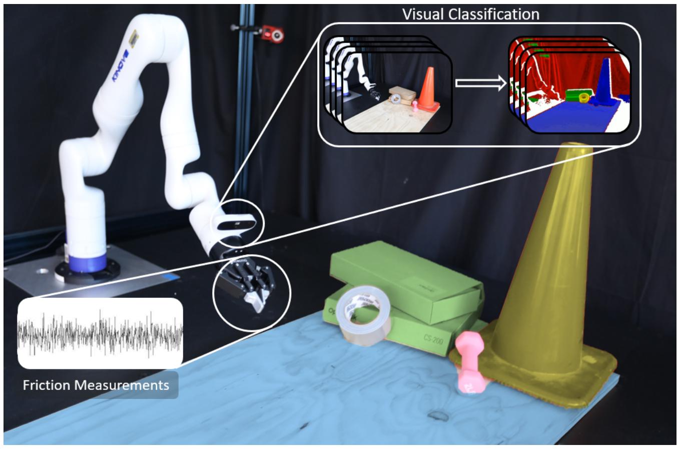

Fig. 1: We condition semantic classifications on their physical properties in order to leverage multiple sensing modalities to increase semantic classification accuracy. Using vision alone, objects in the scene such as the traffic cone and wooden board are incorrectly classified as concrete (navy) and stone (blue), respectively. By taking measurements of the objects' friction coefficients and performing Bayesian inference with the visual semantic classifications we correct classify the objects as plastic and wood (shown in yellow and cyan).

II. Methodology

In the context of learning, semantic classification through vision alone may be an ill-posed problem depending on the scene [4]. The accuracy of these methods rely on the statistics of the dataset used to train the semantic classifier [1]. However, the accuracy of semantic classification may be improved using multi-modal sensing. An example of this is shown in Figure 2.

Multi-modal semantic classification methods have been present since the rise of modern computer vision [5]. These methods often rely on different vision-based modalities such as depth or scale-invariant feature transform (SIFT) [6]. Additionally, many recent multi-modal semantic classification methods are learning-based, where the multi-modal visual features are combined within neural networks to output semantic classifications [7, 8].

While semantic classification methods have seen widespread use in the robotics community there are many robotic tasks, such as footstep or grasp planning, where the underlying semantic properties are more important than the semantic classifications themselves. One example of such properties are affordances, the actionable properties of items, introduced by J.J. Gibson [9]. In our prior work [3], we argue that physical

All authors affiliated with the Robotics Department or the Mechanical Engineering Department at the University of Michigan, 2505 Hayward Street, Ann Arbor, Michigan, USA

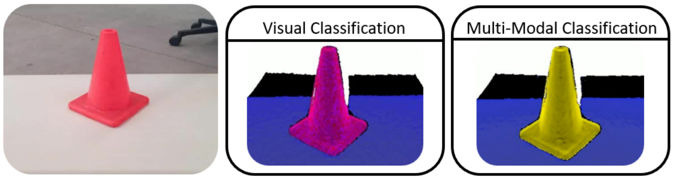

Fig. 2: Semantically classifying the materials present in the image on the left. When only visual information is used, a semantic segmentation network may incorrectly classify the cone as concrete (magenta). In contrast, when we consider friction measurements, our multi-modal framework is able to correctly predict that the traffic cone is plastic (yellow).

properties (e.g., friction, soft contact parameters, stability, etc.) provide the underlying utility of semantic classifications for tasks such as legged locomotion. Our insight was that these properties are conditioned on the semantic classifications. We extend that work here to show that inverse is also true, semantic classifications may be conditioned on their properties.

We leverage this insight to derive a Bayesian inference framework for multi-modal semantic classification and property estimation. In Section II-A we review how properties are conditioned on vision-derived semantic classifications. Next, in Section II-B we extend this line of reasoning and demonstrate how semantic classifications may also be conditioned on their physical properties. We show that this enables us to leverage multi-modal sensing for semantic classification.

A. Recursive Vision-Based Property Estimation

We review our method for recursive semantic classification given vision-based measurements and further demonstrate how to condition properties on these semantic classifications. Next, we introduce the relevant notation.

Let $\boldsymbol{z}$ denote a random vector. The Categorical distribution is a discrete k-dimensional distribution parameterized by $\boldsymbol{\theta} \in [0,1]^k$. The probability mass function of the Categorical distribution represents the probability that sample $\boldsymbol{z}$ belongs to class i, where $i \in \{1, \ldots, k\}$:

$$f(\boldsymbol{z} = i|\boldsymbol{\theta}) = \theta_i. \tag{1}$$

We use (1) to represent the pixel-wise semantic classification measurement output by a semantic classification network. To enforce spatial consistency, we project these pixel-wise measurements onto a map which is static relative to a global frame.

We recursively estimate the categorical likelihood using a Dirichlet distribution for every element (e.g: voxel or mesh element) of this map. The Dirichlet distribution is a continuous k-variate probability distribution parameterized by $\boldsymbol{\alpha} \in \mathbb{R}^k_{\geq 0}$. The probability density function of the Dirichlet distribution is defined as:

$$f(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{j=1}^{k} \alpha_j)}{\sum_{j=1}^{k} \Gamma(\alpha_j)} \prod_{j=1}^{k} \theta_j^{\alpha_j - 1} \tag{2}$$

where

$$\Gamma(\alpha_j) = \int_0^\infty x^{\alpha_j - 1} \exp(-x) dx. \tag{3}$$

The notable property of the Dirichlet distributions is that it is the conjugate prior to the Categorical distribution. This means that the posterior predictive distribution $f(\boldsymbol{\theta}|\mathcal{Z}, \tilde{\boldsymbol{\alpha}})$ given Categorical data $\mathcal{Z}$ remains a Dirichlet Distribution parameterized by $\tilde{\boldsymbol{\alpha}}$ if the prior $f(\boldsymbol{\theta}|\boldsymbol{\alpha})$ is also a Dirichlet distribution. The $\tilde{\boldsymbol{\alpha}}$ parameters are computed recursively as:

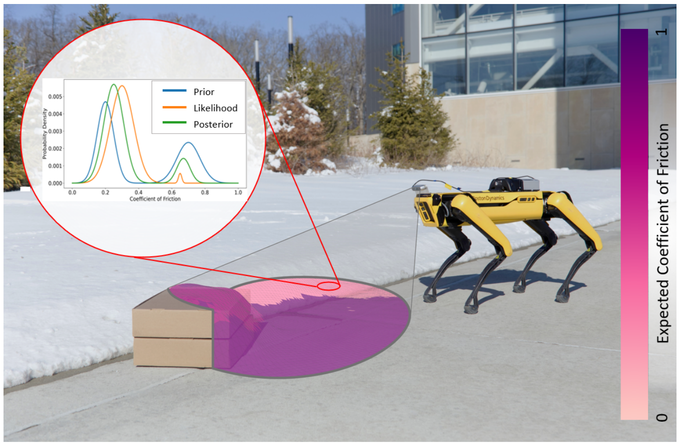

Fig. 3: Our vision-based recursive semantic property estimation algorithm running on a quadrupedal robot outdoors. Semantic classes are estimated recursively by first projecting them onto a spatially-stationary global geometric representation and then tracking class likelihoods using (4). Properties, in this case the coefficient of friction, are conditioned on the semantic class and estimated using (6).

$$\tilde{\alpha}_j = \alpha_j + \sum_{z_i \in \mathcal{Z}} 1\{z_i = j\}, \tag{4}$$

where $1\{z_i = j\}$ is equal to 1 when the expected semantic class of measurement z_i is class j and is zero otherwise. This recursive semantic class update is performed for each element of the spatial representation.

After updating the $\tilde{\boldsymbol{\alpha}}$ parameters the semantic classification likelihood is then computed as:

$$f(z = i|\mathcal{Z}, \tilde{\boldsymbol{\alpha}}) = \frac{\tilde{\alpha}_i}{\sum_{j=1}^{k} \tilde{\alpha}_j}. \tag{5}$$

It was shown in [3] that a uni-modal Gaussian distribution accurately modelled the coefficient of friction for surface material classes. The per-class mean and variance parameters, μ_i and σ_i respectively, were computed using the friction dataset provided in [3]. Given these parameters and the semantic class estimates from (5), the probability distribution for the coefficient of friction ψ is computed as:

$$f(\psi \mid \mathcal{Z}, \boldsymbol{\alpha}) = \sum_{i=1}^{k} \frac{\alpha_i}{\sum_{j=1}^{k} \alpha_j} \mathcal{N}(\mu_i, \sigma_i^2). \tag{6}$$

We demonstrate this recursive property estimation approach in Figures 3 and 4. A trivial modification of (6) allows for other, potentially non-Gaussian, properties to be recursively estimated from vision.

B. Multi-Modal Semantic Classification

In this section we show that semantic classifications may also be conditioned on the semantic properties. Our goal then is to use property measurements, in this case measurements of the coefficient of friction, to jointly compute the semantic classification and property posterior. We once again leverage the theory of conjugate priors to accomplish this.

Let $\Theta = \{\alpha_i, \mu_i, \sigma_i\}_{i=1}^k$ be the set of uncertain parameters for each semantic class introduced in Section II-A. Let $Dir(\cdot)$ and $\mathcal{NG}(\cdot)$ represent the Dirichlet and Normal-Gamma distributions, respectively. We assign as a prior to Θ the Dirichlet Normal-Gamma product distribution:

$$f(\Theta) = Dir(\boldsymbol{\alpha}|\boldsymbol{\beta}) \prod_{i=1}^{k} \mathcal{NG}(\mu_i, \lambda_i | a_i, \kappa_i, b_i, \gamma_i). \tag{7}$$

The parameters are initialized for the prior using the $\boldsymbol{\alpha}$ values computed in Section II-A and the μ_i and σ_i parameters for each surface material found in [3]. Furthermore, we set the initial values of $\boldsymbol{\beta}$, a, b, κ, and γ as

$$\begin{aligned} \boldsymbol{\beta} &= \alpha, \\ a &= \mu, \\ \kappa &= 1, \\ b &= \frac{1}{(C\sigma^2\kappa)^2}, \\ \gamma &= \sqrt{b}/C, \\ C &= 40. \end{aligned} \tag{8}$$

Given a measurement ψ_1 for the coefficient of friction we compute analytical the posterior of Θ using Bayes' theorem as:

$$\begin{aligned} f(\Theta|\psi_1) = \frac{1}{Z} \sum_{j=1}^{k} c_j Dir(\boldsymbol{\alpha}|\tilde{\boldsymbol{\beta}}_j) \mathcal{NG}(\mu_j, \lambda_j | \tilde{a}_j, \tilde{\kappa}_j, \tilde{b}_j, \tilde{\gamma}_j) \cdot \\ \cdot \prod_{j \neq i}^{k} \mathcal{NG}(\mu_i, \lambda_i | a_i, \kappa_i, b_i, \gamma_i). \end{aligned} \tag{9}$$

where

$$\begin{aligned} \tilde{\boldsymbol{\beta}}_j &= \boldsymbol{\beta}_j + 1 \\ \tilde{a_j} &= \frac{\kappa_j a_j + \psi_1}{\kappa_j + 1} \\ \tilde{\kappa_j} &= \kappa_j + 1 \\ \tilde{b_j} &= b_j + \frac{1}{2} \\ \tilde{\gamma_j} &= \gamma_j + \kappa_j \frac{(\psi_1 - a_j)^2}{2(1 + \kappa_j)} \\ c_j &= \sqrt{\frac{\kappa_j}{\tilde{\kappa_j}}} \frac{\Gamma(\tilde{b}_j)}{\Gamma(b_j)} \frac{{\gamma_j}^{(b_j)}}{\tilde{\gamma}_j^{(\tilde{b}_j)}} \end{aligned} \tag{10}$$

From (9) we see that the number of terms in the posterior grows exponentially with the number of measurements. To overcome this challenge, we implement the Bayesian moment matching algorithm [10] to approximate the posterior as a Dirichlet Normal-Gamma product distribution, the same distribution as the prior:

$$f(\Theta|\psi) \approx Dir(\boldsymbol{\alpha}|\hat{\boldsymbol{\beta}}) \prod_{i=1}^{k} \mathcal{NG}(\mu_i, \lambda_i | \hat{a}_i, \hat{\kappa}_i, \hat{b}_i, \hat{\gamma}_i) \tag{11}$$

where $\hat{\boldsymbol{\beta}}$, $\hat{a}_i$, $\hat{\kappa}_i$, $\hat{b}_i$, and $\hat{\gamma}_i$ are computed using the Bayesian moment matching algorithm discussed next.

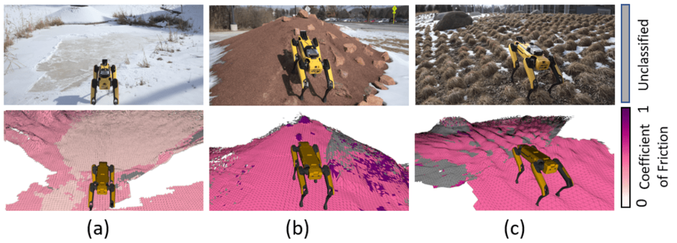

Fig. 4: Our recursive vision-based property estimation algorithm is demonstrated on three outdoor scenes: near a frozen pond, on a sandy hill, and surrounded by low vegetation.

The distribution in (11) has six moments, or six sufficient statistics, which uniquely determine the distribution. Thus, to approximate the posterior (9) using (11), we match the moments of (11) with the first six moments of (9).

We first compute the following quantities:

$$\begin{aligned} \mathbb{E}[\mu_i] &= \sum_{j=1}^{k} \frac{c_j}{\sum_{i=1}^k c_i} (a_i) \\ \mathbb{E}[\lambda_i] &= \sum_{j=1}^{k} \frac{c_j}{\sum_{i=1}^k c_i} \left(\frac{b_i}{\gamma_i} \right) \\ \mathbb{E}[\lambda_i^2] &= \sum_{j=1}^{k} \frac{c_j}{\sum_{i=1}^k c_i} \left(b_i \frac{(b_i + 1)}{\gamma_i^2} \right) \\ \mathbb{E}[\mu_i \lambda_i^2] &= \sum_{j=1}^{k} \frac{c_j}{\sum_{i=1}^k c_i} \left(\frac{1}{\kappa_i} + a_i^2 \frac{b_i}{\gamma_i} \right) \\ \mathbb{E}[\boldsymbol{\alpha}_i] &= \sum_{j=1}^{k} \frac{c_j}{\sum_{i=1}^k c_i} \left(\frac{\boldsymbol{\beta}_i}{\sum_{i=1}^k \boldsymbol{\beta}_i} \right) \\ \mathbb{E}[\boldsymbol{\alpha}_i^2] &= \sum_{j=1}^{k} \frac{c_j}{\sum_{i=1}^k c_i} \left(\frac{\boldsymbol{\beta}_i}{(\sum_{i=1}^k \boldsymbol{\beta}_i)} \frac{\boldsymbol{\beta}_i + 1}{(\sum_{i=1}^k \boldsymbol{\beta}_i + 1)} \right) \end{aligned} \tag{12}$$

We then use these equations to compute the sufficient statistics of (11):

$$\begin{aligned} \hat{a}_i &= \mathbb{E}[\mu_i] \\ \hat{\kappa}_i &= \frac{1}{\mathbb{E}[\mu_i \lambda_i^2] - \mathbb{E}[\mu_i]^2 \mathbb{E}[\lambda_i]} \\ \hat{b}_i &= \frac{\mathbb{E}[\lambda_i]^2}{\mathbb{E}[\lambda_i^2] - \mathbb{E}[\lambda_i]^2} \\ \hat{\gamma}_i &= \frac{\mathbb{E}[\lambda_i]}{\mathbb{E}[\lambda_i^2] - \mathbb{E}[\lambda_i]^2} \\ \hat{\boldsymbol{\beta}}_i &= \mathbb{E}[\boldsymbol{\alpha}_i] \frac{\mathbb{E}[\boldsymbol{\alpha}_i] - \mathbb{E}[\boldsymbol{\alpha}_i^2]}{\mathbb{E}[\boldsymbol{\alpha}_i^2] - \mathbb{E}[\boldsymbol{\alpha}_i]^2} \end{aligned} \tag{13}$$

When multiple property measurements are given, the posterior is updated sequentially.

What (11) shows is that by incorporating information from measurements ψ we update parameters $\boldsymbol{\alpha}$, representing class likelihood, as well as the class-dependent property parameters, μ_i and σ_i. Intuitively, this step updates the class likelihood parameters, $\boldsymbol{\alpha}$, such that the class likelihood is similar to the class whose properties agree with the measurement.

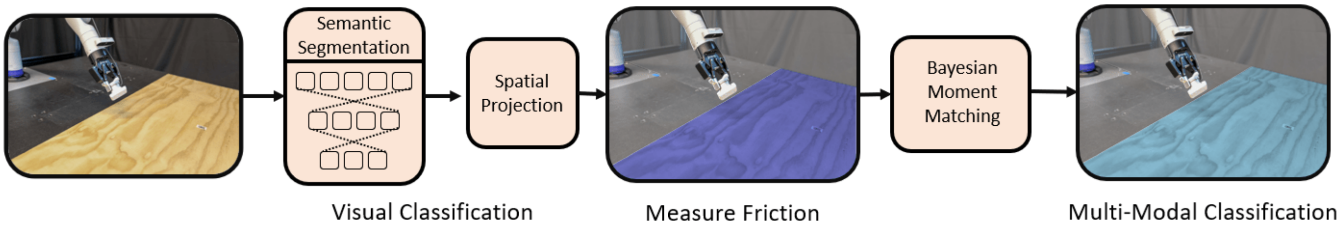

Fig. 5: The pipeline for our approach is shown. Vision is used to recursively estimate the semantic class of the surface following [3]. First, RGB-D images are passed through a semantic segmentation network and then projected onto a map. Using vision alone, the surface is incorrectly classified as stone (blue). Friction measurements are then used to update these semantic estimates via the Bayesian Moment Matching algorithm using (9). After the update, the surface is correctly classified as wood (cyan).

III. Results

We demonstrate our approach in several scenarios. First, we show qualitative results for recursive, vision-based property estimation reviewed in Section II-A. Additionally, we show two experiments where we use friction measurements to correct erroneous semantic classifications.

A. Implementation

For experiments shown in Figures 3 and 4 we use a 2.5 elevation map as the spatial representation. Our recursive vision-based property estimation algorithm is implemented for this elevation map in C++ and includes a Robot Operating System (ROS) interface[1]. Our implementation features noise models for the Realsense RGB-D camera and an interface to include additional sensor noise models. For the remaining experiments we implement a variant of the Kinect Fusion [11] signed distance field spatial representation which can also store semantic segmentation labels. To compute semantic segmentations from RGB images we combine the FastSAM [12] segmentation network and the SegFormer [13] semantic classification network trained on the Apple Material Segmentation Dataset [14] which is able to run at 10 frames per second. We evaluated our method on a laptop with a 3.1GHz Ryzen 3600 processor, 32GB of RAM and an Nvidia RTX 2080 Ti GPU.

B. Recursive Vision-Based Property Estimation

Figure 3 illustrates how new semantic classifications from vision are used to update the probability density function of the coefficient of friction. Figure 4 demonstrates recursive property estimation near a frozen pond, a sandy hill, and in a region with low vegetation. Note, the property estimates are shown using a single color representing the mean coefficient of friction of the most likely terrain class, however each triangular element of mesh estimates the probability density function for the coefficient of friction as per (6).

C. Multi-Modal Semantic Classification

In Figures 1 and 2 we demonstrate two scenarios where vision-based semantic classification fails and property measurements are used to correct these erroneous classifications. Semantic property priors for all material classes used in these experiments are taken from our previous work [3].

In Figure 2, our semantic classification network incorrectly estimates the traffic cone as the concrete material class. After friction measurements are taken and the posterior is computed as per (9), the material is correctly predicted to be plastic.

In Figure 1, a Kinova robotic arm observes a wooden panel which is incorrectly classified as stone. Again, using friction measurements we are able to correctly predict the material as wood.

Both of these examples show how our approach is able to overcome the deficiencies of the semantic segmentation network using only a small, easy-to-collect dataset of material friction coefficients. The friction coefficient priors are taken from our previous work [3]. Lastly, Figure 5 demonstrates the pipeline of our approach.

IV. Conclusion

We present our approach for multi-modal semantic segmentation. By conditioning semantic class on physical properties we are able to leverage multiple sensing modalities in combination with vision to more accurately predict semantic classifications. In contrast to exist multi-modal classification methods, we use Bayesian inference to update semantic classifications without the use of neural networks apart from image segmentations. We demonstrated our approach on several real-world scenes where friction measurements are used to correct erroneous semantic classifications.

References

[1] B. Emek Soylu, *et al.*, “Deep-learning-based approaches for semantic segmentation of natural scene images: A review,” *Electronics*, vol. 12, no. 12, 2023. [Online]. Available: https://www.mdpi.com/2079-9292/12/12/2730

[2] T.-N. Le, *et al.*, “Camouflaged instance segmentation in-the-wild: Dataset, method, and benchmark suite,” *IEEE Transactions on Image Processing*, vol. 31, 2021.

[3] 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, 2022.

[4] P. Yee and S. S. Haykin, “Pattern classification as an ill-posed, inverse problem: a regularization approach,” *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 597–600 vol.1, 1993. [Online]. Available: https://api.semanticscholar.org/CorpusID:2711442

[1] https://github.com/roahmlab/sel_map

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

[6] 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.

[7] 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.

[8] 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.

[9] J. J. Gibson, “The theory of affordances,” *Hilldale, USA*, vol. 1, no. 2, 1977.

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

[11] S. Izadi, *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.

[12] X. Zhao, *et al.*, “Fast segment anything,” *arXiv preprint arXiv:2306.12156*, 2023.

[13] E. Xie, *et al.*, “Segformer: Simple and efficient design for semantic segmentation with transformers,” *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[14] P. Upchurch and R. Niu, “A dense material segmentation dataset for indoor and outdoor scene parsing,” in *European Conference on Computer Vision*. Springer, 2022.