Robotics: Science and Systems 2020
Corvalis, Oregon, USA, July 12-16, 2020

Efficient Parametric Multi-Fidelity Surface Mapping

Aditya Dhawale and Nathan Michael
The Robotics Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania, 15213
Email: {adityand, nmichael}@cmu.edu

Abstract—**State-of-the-art dense mapping approaches cannot be deployed on Size, Weight, and Power (SWaP) constrained platforms because of their large memory and compute requirements. In this paper, we present an accurate, and efficient approach to dense multi-fidelity 3D mapping using Gaussian distributions as volumetric primitives. The proposed mapping approach supports both high fidelity dense surface reconstruction and lower fidelity volumetric environment representation for fundamental robotics applications. We exploit the inherent working characteristics of an off-the-shelf depth sensor and approximate the distribution of approximately planar points using Gaussian distributions. Explicit modeling of the sensor noise characteristics enable us to incrementally update the map representation in real-time with high accuracy. We present the advantages of our proposed map representation over other well known state-of-the-art representations by highlighting its superior performance in terms of reconstruction accuracy, completeness and map compression properties via quantitative and qualitative metrics.**

I. INTRODUCTION

A Size, Weight, and Power (SWaP) constrained autonomous system deployed in real-world environments to enable infrastructure inspection, exploration, search and rescue, must solve various challenging problems such as consistent mapping, safe planning and localization in real-time. Spitzer et al. [23] describe a semi-autonomous aerial robotic system exploring unknown environments that performs state-estimation, local mapping and collision avoidance in real-time at high speeds using multiple independent map representations. A disjoint perceptual system increases the memory and computational burden on a robot thus affecting its accuracy and speed of performance. In this paper, we present a multi-fidelity probabilistic mapping strategy using Gaussian distributions as structure primitives. The hierarchical structure of our proposed map representation enables the unification of a disjoint autonomous system and reduces the computational burden on a SWaP constrained system.

RGB-D sensors operating in real-world environments observe 3D data along the surfaces of objects in the scene. Object surfaces can be approximated as locally planar and reconstructed using planar surface elements. Our proposed mapping framework uses Gaussian distributions as structure primitives that capture the local structural relationships in the observed data and create a continuous representation of the objects in the scene at required fidelity. We thus exploit the local regularity and the considerably reduced storage complexity of Gaussian distributions to represent the mapping as a model fitting problem.

In this paper, we present a unified hierarchical mapping framework that generates a highly accurate representation for dense surface mapping and more succinct representations at higher hierarchical levels, that can be utilized for pose estimation [4, 25] and global path planning [5, 26]. Further, our proposed framework explicitly models the uncertainty of noisy depth information to enable more accurate and concise 3D reconstruction of the scene observed using Commercially available Off-The-Shelf (COTS) RGB-D sensors.

We represent a depth camera observations as 3D parametric uncertain distributions and use them to construct an uncertainty aware map representation. This map representation is refined with a stream of sensor measurements to represent accurately the surfaces of objects in the scene. A Commercial off-the-shelf (COTS) structured light depth sensor projects a structured pattern of IR light onto its surroundings and measures the distance of all the surfaces in its field-of-view. We exploit the understanding that the ordered data obtained from such sensors resides along the surfaces in the scene within the ambient space $\mathbb{R}^3$. Our proposed approach replaces the most computationally expensive part of fitting GMMs, EM [6, 24], with a simpler pixel space search-based "Region Growing" technique [19], by constraining each separate component of a GMM to represent different sections of the surfaces in the scene. Our main contributions are:

- A novel pixel space search-based surface mapping technique using Gaussian distributions as surface elements (Sec. III-B)
- A probabilistic map update strategy using sensor noise modeling to compute more accurate correspondences between sensor observations and the map (Sec. III-C)
- A hierarchical multi-fidelity mapping strategy that enables consistent high fidelity map updates(Sec. III-E)

We demonstrate (Sec. V) that our approach is more memory efficient in terms of memory than state-of-the-art mapping techniques while retaining the most information about the map with high precision and low reconstruction error.

II. RELATED WORK

The choice of map representation used for a perceptual framework on an autonomous system is driven by the application and the computational and memory resources available on-board. For path planning and 3D navigation, a majority of mobile robots employ voxel grids [7] as the map representation. Voxel grids assume conditional independence of each voxel

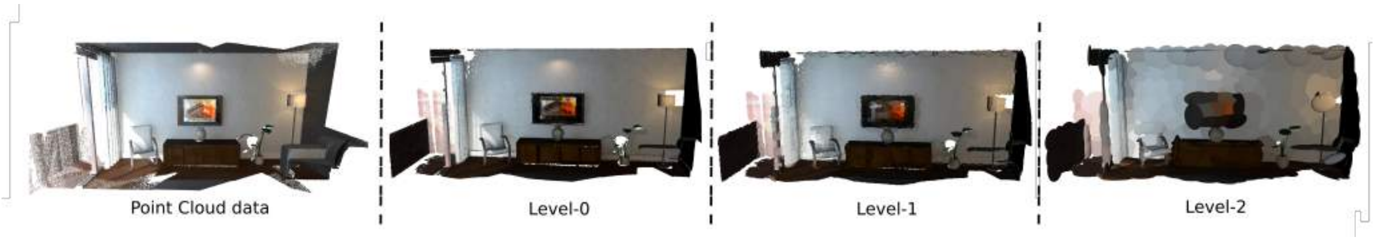

Fig. 1: *Left To Right:* As a sequence of 3D point cloud data is observed, a Gaussian distribution based consistent map of the world ${}^{w}\mathbf{\Theta}^{0}$ is initialized at high fidelity; These Gaussian distributions are combined together using a pixel-space similarity based surface growing algorithm to generate lower fidelity representations of the world at hierarchical level-1 and level-2. The 3σ bound ellipsoids of each Gaussian distribution are visualized colored by the mean color of the 3D points approximated by each distribution.

from its neighboring voxels that fails to capture the spatial dependency that the real-world data exhibits. The fidelity at which the environment can be represented is further limited by the resolution of the voxel grid. Hornung et al. [9] proposed an octree based solution to reduce the memory complexity of voxel grids by adaptively allocating more memory resources in complex regions of the map.

Dense Simultaneous Localization and Mapping approaches [21, 28, 10] generate an accurate 3D reconstruction of the world by reconstructing the surfaces in the scene. Schops et al. [21], Whelan et al. [28] approximate the locally planar spread of points using small disk like surface elements (surfels). Izadi et al. [10] on the other hand, use a Truncated Signed Distance Field (TSDF) to implicitly represent a surface. These dense mapping algorithms are extremely computationally expensive, and have a large memory footprint and therefore pose a major challenge in deploying on SWaP constrained robots.

Several approaches have been proposed to reduce the computational and storage complexity of surface mapping [27, 14, 15]. The former represents large planar regions in the scene using a parametric planar representation. However, the strong assumption of planarity over large regions fails in complex and outdoor environments. The latter two approaches represent surfaces using Gaussian processes. Such non-parametric approaches to mapping are capable of representing the surface information at high fidelity. However, they require large amount of data and fitting GP parameters is computationally complex. Therefore training a GP to dense point cloud data provided by RGBD sensors in real-time is challenging. Similar in vein to our proposed approach, Pizzoli et al. [18] propose a probabilistic approach to generate depth maps using monocular color information by explicitly modeling uncertain data. However, this approach is reliant on availability of texture in the color information which is often not available in indoor or outdoor environments.

A novel map representation has been proposed by Magnusson et al. [16], Srivastava and Michael [24], Eckart et al. [6] that use succinct Gaussian Mixture Models (GMMs) to reduce the memory complexity of map representations while retaining high fidelity of representation. Normal Distributions Transform (NDT) maps proposed by Magnusson et al. [16], unlike the proposed framework, fit Gaussian distributions over each discretized 3D cell thus leading to lower fidelity of representation at cell boundaries. Eckart et al. [6], Srivastava and Michael [24] present two hierarchical 3D mapping approaches that fit a consistent GMM over the 3D structure data. However, the Expectation Maximization routine used in these approaches requires the information about model complexity a-priori, is extremely computationally expensive, and, is highly sensitive to parameter initialization [11].

III. APPROACH

This section describes our proposed mapping framework. Instead of employing an expensive EM routine, we use a pixel space search-based “Region Growing” technique to fit Gaussian distributions to planar surfaces in the scene thus providing a more accurate measure of reconstruction accuracy. It also enables us to achieve orders of magnitude of computational savings, thus making it feasible to learn accurate Gaussian distribution based representations of the sensor data in real-time. Fig. 2 illustrates the system overview of our algorithm.

The mapping process can be summarized into the following steps:

1) Initialize a high resolution surface representation of the environment (Sec. III-B)
2) Compute a correspondence map that denotes the point-to-distribution correspondence for each point in a new depth measurement (Sec III-C1)
3) Fuse incrementally observed noisy sensor data into the global map representation using correspondence map and Gaussian merging (Sec. III-C)
4) Map refinement using Gaussian splitting and outlier removal (Sec. III-D)
5) Hierarchical map update (Sec. III-E)

A. GAUSSIAN DISTRIBUTION

A Gaussian distribution ${}^{w}\boldsymbol{\theta}$ in a d-dimensional space, defined in a coordinate frame w, is a probability distribution function that is parameterized by its mean and covariance:

$${}^{w}\boldsymbol{\theta}_j := (\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$

A set of points ${}^{w}\mathbf{X} := \{{}^{w}\mathbf{x}_0, {}^{w}\mathbf{x}_1, \ldots, {}^{w}\mathbf{x}_N\}$ can be approximated with a Gaussian distribution ${}^{w}\boldsymbol{\theta}_i$ as:

$$\boldsymbol{\mu}_j = \frac{\sum_i^N {}^{w}\mathbf{x}_i}{N}, \boldsymbol{\Sigma}_j = \frac{\sum_i^N {}^{w}\mathbf{x}_i {}^{w}\mathbf{x}_i^T}{N} - \boldsymbol{\mu}_j \boldsymbol{\mu}_j^T \qquad (1)$$

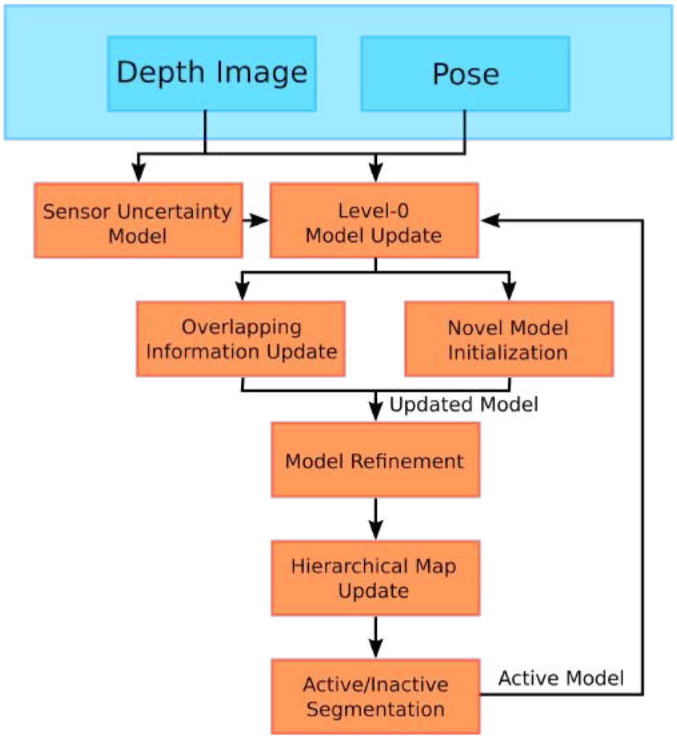

Fig. 2: System Overview: The algorithm operates on depth image streams and corresponding sensor poses. As sequential sensor measurements are observed, an accurate map of the world ${}^{w}\boldsymbol{\Theta}^{0}$ is updated with novel and overlapping sensor information. Noisy Gaussian distributions are removed from the map and the higher hierarchical levels ${}^{w}\boldsymbol{\Theta}^{1}, {}^{w}\boldsymbol{\Theta}^{2}$ are updated using the refined ${}^{w}\boldsymbol{\Theta}^{0}$. The Gaussian distributions outside the FOV of the sensor are labeled inactive to reduce the computational complexity of updating large scale maps.

B. SURFACE MODEL INITIALIZATION

We initialize a Gaussian distribution based representation of the environment at hierarchical level-0 as the first sensor measurement is observed at time $t = 0$ along with its corresponding ground truth pose in a coordinate frame w. This representation is formulated as an ordered set of Gaussian distributions in a global coordinate frame w, ${}^{w}\boldsymbol{\Theta}^{0}_{0} = \left\{{}^{w}\boldsymbol{\theta}^{0}_{0}, {}^{w}\boldsymbol{\theta}^{0}_{1}, \dots {}^{w}\boldsymbol{\theta}^{0}_{n}\right\}$. The covariance $\boldsymbol{\Sigma}_i$ of each ${}^{w}\boldsymbol{\theta}^{0}_{i}$ signifies the spread of points that it is fit on. Depth measurements obtained from COTS structured light sensors are spread along sub-manifolds (surfaces of the objects in the scene) in the ambient $\mathbb{R}^3$ space. A ${}^{w}\boldsymbol{\theta}^{0}_{i}$ that best represents a subset of this observed data corresponds to a planar patch of the spread of points along a surface. The smallest eigenvalue $\lambda_{i,0}$ of $\boldsymbol{\Sigma}_i$ represents the variance of points along the direction with the least data variation i.e., the normal to the surface that ${}^{w}\boldsymbol{\theta}^{0}_{i}$ is fit on. The corresponding eigenvector $\mathbf{u}_{i,0}$ represents the direction of this normal. Given an image $\mathcal{I}_0$ of size $V \times U$, we use this understanding to fit the best Gaussian distributions ${}^{w}\boldsymbol{\theta}^{0}_{i}$ to small $b_0 \times b_0$ pixel patches of $\mathcal{I}_0$ that represent planar spread of points.

Assuming pinhole camera geometry, a depth measurement at pixel location (v, u), $\mathcal{I}_0(v, u)$, can be back projected in to $\mathbb{R}^3$ space using inverse projective function

$$\mathbf{x}_{v,u} = \boldsymbol{\Pi}^{-1}(\mathcal{I}_0(v, u), v, u) \tag{2}$$

Since, this is a linear transformation, the points that are proximate in the 3D space, are also proximate in pixel space. We exploit this property of projective pinhole geometry and implement a version of the "Region Growing" algorithm proposed by Poppinga et al. [19] that operates in the image space using uncertain data. The "Region Growing" algorithm can be described as follows:

In each image patch of size $b_0 \times b_0$ pixels, we

1) Select a random seed pixel (v_s, u_s) and compute the 3D point $\mathbf{x}_{v_s,u_s} = \boldsymbol{\Pi}^{-1}(\mathcal{I}(v_s, u_s), v_s, u_s)$
2) Search for candidate 3D points $\mathbf{x}_{v_c,u_c}$ in the pixel neighborhood of $\mathbf{x}_{v_s,u_s}$ that lie within α_n distance of $\mathbf{x}_{v_s,u_s}$ and initialize a Gaussian distribution ${}^{w}\boldsymbol{\theta}^{0}_{i}$ using this set of points as defined by Eq. 1
3) For candidate points $\mathbf{x}_{v_c,u_c}$ that are inside α_n distance, search for its neighbor points $\mathbf{x}_{v_k,u_k}$ within α_n distance, such that the smallest eigenvalue λ_0 of the covariance $\boldsymbol{\Sigma}_i$ of a Gaussian distribution ${}^{w}\boldsymbol{\theta}^{0}_{i}$ updated with point $\mathbf{x}_{v_k,u_k}$ is less than $\alpha^2_{0,\lambda}$ and the largest eigenvalue is less than $\alpha_{0,\text{len}}$
4) Continue until all points in the image patch are processed

The mean and covariance of a Gaussian distribution completely represent sufficient information about the points that it was fit on. Therefore, for incrementally updating a Gaussian distribution ${}^{w}\boldsymbol{\theta}^{0}_{j}$ with a point ${}^{w}\mathbf{x}_i$ we do not need to preserve the history of points used to fit ${}^{w}\boldsymbol{\theta}^{0}_{j}$:

$${}^{w}\boldsymbol{\theta}^{0'}_{j} := \left\{\boldsymbol{\mu}^{0'}_{j}, \boldsymbol{\Sigma}^{0'}_{j}, N'_j\right\} \tag{3}$$

$$N'_j = N_j + 1 \tag{4}$$

$$\boldsymbol{\mu}^{0'}_{j} = \frac{N_j \boldsymbol{\mu}^{0}_{j} + {}^{w}\mathbf{x}_i}{N'_j} \tag{5}$$

$$\boldsymbol{\Sigma}^{0'}_{j} = \frac{N_j\left(\boldsymbol{\Sigma}^{0}_{j} + \boldsymbol{\mu}^{0}_{j}\boldsymbol{\mu}^{0\,T}_{j}\right) + {}^{w}\mathbf{x}{}^{w}\mathbf{x}^{T}}{N'_j} - \boldsymbol{\mu}^{0'}_{j}\boldsymbol{\mu}^{0'T}_{j} \tag{6}$$

For each $b_0 \times b_0$ image patch, we select the Gaussian distribution that represents the largest planar region in that patch. Thus, our initial model at hierarchical level-0, ${}^{w}\boldsymbol{\Theta}^{0}_{0}$, consists of $\frac{V}{b_0} \times \frac{U}{b_0}$ Gaussian distributions that represent planar patches in the observed scene. The covariance matrix of each Gaussian distribution is regularized by adding a small ($1e-6$) value, to avoid numerical inconsistencies caused by noisy sensor data and to ensure that each covariance matrix is Positive Semi-Definite (PSD). Fig 1 provides an example of a model $\boldsymbol{\Theta}$ fit at multiple hierarchical levels on a set of input scans.

C. INCREMENTAL MODEL UPDATES

As sequential sensor measurements are obtained, we can refine the current model estimate at the hierarchical level-0, ${}^{w}\boldsymbol{\Theta}^{0}_{t-1}$, of the scene using this incoming stream of structure information. Sequential sensor measurements obtained from a sensor moving in the world, contain some novel information about the scene and some redundant information that has already been observed. However, sensor measurements obtained from a depth sensor are often noisy and therefore unreliable. We use the redundant structure data to improve the estimate of the map that is already observed using a weighted update and fit new Gaussian distributions to novel information. An additional uncertainty covariance $\mathbf{\Sigma}_k^{unc}$ is added to the Gaussian distributions to represent the average uncertainty of the points used to fit each Gaussian distribution, ${}^w\boldsymbol{\theta}_{i-1}^k := \{\boldsymbol{\mu}_k, \mathbf{\Sigma}_k, \mathbf{\Sigma}_k^{\text{unc}}, N_k\}$.

Given a sensor measurement ${}^w\mathbf{x}_{v,u}$ at a pixel location (v, u) in image $\mathcal{I}_t$ with Gaussian uncertainty ${}^w\mathbf{\Sigma}_{v,u}^{\text{unc}}$, we check if this point has already been represented by a distribution in ${}^w\mathbf{\Theta}_{t-1}^0$. If $\left\{{}^w\mathbf{x}_{v,u}, {}^w\mathbf{\Sigma}_{v,u}^{\text{unc}}\right\}$ has a similarity measure higher than α_{conf} with ${}^w\hat{\boldsymbol{\theta}}_k^0 := \left\{\boldsymbol{\mu}_k, \mathbf{\Sigma}_k + \mathbf{\Sigma}_{v,u}^{\text{unc}} + \mathbf{\Sigma}_k^{\text{unc}}\right\}$ then it is considered to be partially represented by ${}^w\boldsymbol{\theta}_k^0$ and therefore, is classified as non-novel. Points ${}^w\mathbf{x}_{k,l}$ that do not lie within the confidence interval of any of the distributions in ${}^w\mathbf{\Theta}_{i-1}^0$ are labeled as novel points. "Bhattacharyya Coefficient" [13] is used as a measure of overlap or similarity between two Gaussian distributions.

The novel information obtained in $\mathcal{I}_t$ is used to initialize the new Gaussian distributions ${}^w\boldsymbol{\theta}_k^0$ at hierarchical level-0 that represent the newly observed surfaces in the scene.

1) CORRESPONDENCE MAP: In order to update the global map of the world ${}^w\mathbf{\Theta}_{t-1}^0$ with the correct overlapping 3D point ${}^w\mathbf{x}_{v,u}$, we must check if the uncertain distribution of ${}^w\mathbf{x}_{v,u}$ overlaps with each uncertain distribution in ${}^w\mathbf{\Theta}_{t-1}^0$. As the number of components in ${}^w\mathbf{\Theta}_{t-1}^0$ increases, the search space for finding the correspondence between a point $\mathbf{x}_{v,u} \in \mathcal{I}_t$ and distribution ${}^w\boldsymbol{\theta}_k^{t-1} \in {}^w\mathbf{\Theta}_{t-1}^0$ also increases. To make the search real-time, we exploit the geometric properties of a Gaussian distribution and the linearity of pinhole projection. Gaussian distributions have a infinite support space in 3D. However, the probability distribution tapers off exponentially farther away from the mean. Equipotential contours of a Gaussian distribution can geometrically be represented as ellipsoids in $\mathbb{R}^3$ space. The pinhole projection of a 3D ellipsoid is an ellipse in a 2D image plane [4]. Only the Gaussian distributions corresponding to ellipses that project at a given pixel location (v, u) on the image plane are potentially proximate to the 3D point $\mathbf{x}_{v,u}$ and contribute to its likelihood. We implement a depth buffer in OpenGL to compute projectively correct point to distribution correspondences as shown in Fig. 3. A 3σ ellipsoid corresponding to each Gaussian distribution is used for projection. Finally, to verify if the projective correspondence is correct, we check if the uncertain point $\mathbf{x}_{v,u}$ distribution overlaps with the corresponding Gaussian distribution ${}^w\boldsymbol{\theta}_k^0$ that was projected at the same pixel location.

2) GAUSSIAN DISTRIBUTION UPDATE: If an uncertain point $\left\{{}^w\mathbf{x}_{v,u}, {}^w\mathbf{\Sigma}_{v,u}^{\text{unc}}\right\}$ has been previously partially observed by a distribution ${}^w\boldsymbol{\theta}_k^0$, ${}^w\boldsymbol{\theta}_k^0$ can be interpreted as the prior probability of a 3D point being sampled around ${}^w\mathbf{x}_{v,u}$. We can therefore compute the posterior probability as a product of two Gaussian distributions [2], given by

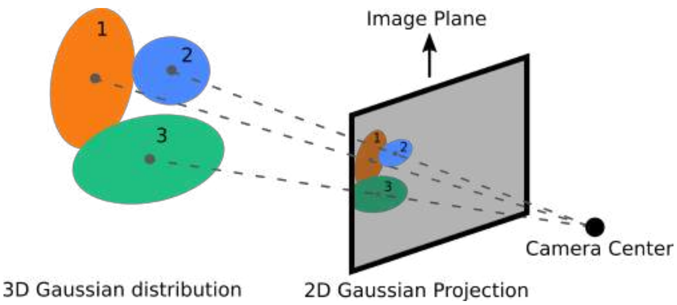

Fig. 3: Illustrative example of projective correspondence. 3σ ellipsoids of 3D Gaussian distributions are projected in the image space that represent 2D ellipses. The 2D ellipses store the index of the 3D Gaussian distribution at each projected pixel location. Correspondence between a point and a Gaussian distribution is computed by querying the index at the corresponding pixel location.

$$\begin{aligned} \hat{\mathbf{\Sigma}}_j^{\text{unc}} &= \left((\mathbf{\Sigma}_k)^{-1} + \left(\mathbf{\Sigma}_j^{\text{unc}}\right)^{-1}\right)^{-1} \\ \hat{\mathbf{x}}_j &= \hat{\mathbf{\Sigma}}_j^{\text{unc}}\mathbf{\Sigma}_k^{-1}\mathbf{x}_j + \hat{\mathbf{\Sigma}}_j^{\text{unc}}\mathbf{\Sigma}_j^{\text{unc}}\boldsymbol{\mu}_{i-1}^k \end{aligned} \tag{7}$$

The posterior distribution ${}^w\hat{\mathbf{X}}_{v,u} := \left\{{}^w\hat{\mathbf{x}}_{v,u}, {}^w\hat{\mathbf{\Sigma}}_{v,u}^{\text{unc}}\right\}$ defines the most likely estimate of the partially observed noisy point. We can add this point to the current estimate of ${}^w\boldsymbol{\theta}_{i-1}^k$ and refine its parameters.
If the sensor uncertainty model is perfectly known, the incremental update defined in Eq. 7 provides an accurate reconstruction of the structure in the scene with minimal sensor information.

D. MODEL REFINEMENT

Sensor measurements obtained from RGB-D sensors often contain spurious depth measurements due to reflective objects in the scene, noisy particles or just random noise. If a distribution ${}^w\boldsymbol{\theta}_k^0$ is fit to such noisy data, the proposed projective correspondence computation pipeline fails as these components occlude the true map distributions. However, due to the spurious nature of such measurements, sequential sensor measurements do not provide overlapping evidence for the noisy distributions, ${}^w\boldsymbol{\theta}_k^0$. The pixel locations where ${}^w\boldsymbol{\theta}_k^0$ is projected are rejected as candidate correspondences by our correspondence refinement step. If the rendered depth measurements at these pixel locations are less than the observed depth measurements, the number of points represented by ${}^w\boldsymbol{\theta}_k^0$, N_k is set to $N_k - 1$. Distributions ${}^w\boldsymbol{\theta}_k^0$ that do not have sufficient evidence $N_k < N$ are eventually discarded thus eliminating spurious and noisy distributions from the hierarchical level-0 map ${}^w\mathbf{\Theta}_t^0$.

E. HIERARCHICAL MAPPING

The model reconstruction strategy described in Sec. III-B is inherently independent for each image patch and therefore, easily extendible to a hierarchical framework. Similar to the level-0 pixel space region-growing, we employ a "Region Growing" approach modified to use Gaussian distributions, ${}^w\boldsymbol{\theta}_t$ as the primitives instead of points ${}^w\mathbf{x}$. A 3-level hierarchy is defined in image space by dividing the image into larger image patches recursively as shown in Fig. 4. We divide the image into image patches of size $b_1 \times b_1$ at hierarchical level-1 and $b_2 \times b_2$ at hierarchical level-2 such that $b_2 \geq b_1 \geq b_0$. Given the map fitted at hierarchical level-0, ${}^{w}\mathbf{\Theta}_t^0$, a Gaussian distribution based "Region Growing" is performed in a image patch of size $b_1 \times b_1$, using "Bhattacharyya Coefficient" as a similarity measure with thresholds on the maximum thickness of a Gaussian distribution $\alpha_{1,\lambda}$ and the largest spread of a Gaussian distribution $\alpha_{1,\text{len}}$.

Once all ${}^{w}\mathbf{\Theta}_t^1$ distributions are computed, the same process is repeated in a larger image patch of size $b_2 \times b_2$ with thresholds $\alpha_{2,\lambda}$ and $\alpha_{2,\text{len}}$. Finally, every distribution ${}^{w}\boldsymbol{\theta}_k^2$ absorbs one or more distributions $\boldsymbol{\theta}_k^1$ which absorbs one or more $\boldsymbol{\theta}_t^0$. Due to the distributive nature of our image patch splitting, each distribution in hierarchical level-$l-1$ can only have one parent distribution in hierarchical level-l that it is encompassed by. An illustrative example of a hierarchical map learned on a single depth measurement is shown in Fig. 4.

As the model ${}^{w}\mathbf{\Theta}_t^0$ is incrementally updated, these updates are propagated up the hierarchical tree structure at each time step and therefore the entire hierarchical map is updated with the most recent information.

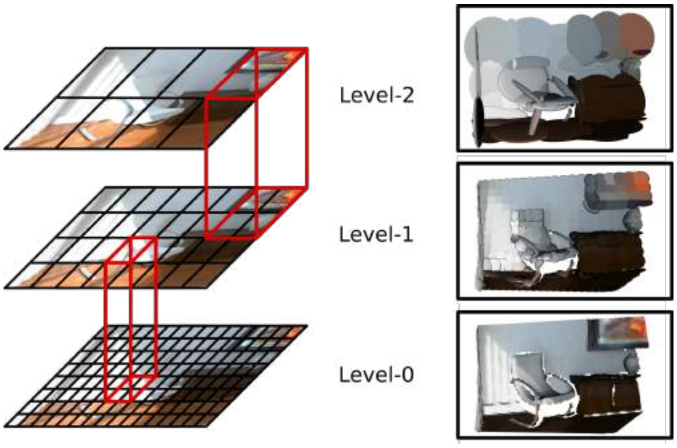

Fig. 4: Illustrative example of the proposed hierarchical mapping strategy. Given a depth image $\mathcal{I}_t$ at time t our hierarchical approach splits the image into fine grids and fits a high resolution map ${}^{w}\mathbf{\Theta}_t^0$ according to Sec. III-B on image patches of size $b_0 \times b_0$. Distributions in this model that lie on a larger image patch of size $b_1 \times b_1$ are combined to create ${}^{w}\mathbf{\Theta}_t^1$ and similarly distributions in ${}^{w}\mathbf{\Theta}_t^1$ that lie on an even larger image patch of size $b_2 \times b_2$ are further combined to create ${}^{w}\mathbf{\Theta}_t^2$

IV. Implementation

A. ACTIVE-INACTIVE MAPPING

RGB-D sensors like the Kinect have a limited FOV. As a stream of sensor measurements is obtained, only a part of the observed map is currently in the sensor FOV. Therefore, we can subselect the number of distributions that will be affected by the current sensor measurement to only the distributions that lie within the FOV of the sensor. The model at hierarchical level-2 ${}^{w}\mathbf{\Theta}_t^2$ has a very small number of distributions and can be used to efficiently check if the individual distributions ${}^{w}\boldsymbol{\theta}_k^2$ lie within the FOV of the sensor. All ${}^{w}\mathbf{\Theta}_t^0$ distributions whose corresponding $\mathbf{\Theta}_t^2$ distribution lies outside the current FOV of the sensor, are labeled inactive, and the rest of the map is active.

B. SENSOR UNCERTAINTY MODEL

We use the sensor noise model presented in [20] where the uncertainty of the depth measurements obtained from an RGB-D kinect sensor is represented as a Gaussian distribution along each ray, as

$$\mathbf{\Sigma}^{\mathbf{x}} = \mathbf{J}_{\mathbf{x}} \text{diag}\left(\mathbf{\Sigma}_p, \mathbf{\Sigma}_p, \sigma_z^2\right) \mathbf{J}_{\mathbf{x}}^T \tag{8}$$

where (σ_z) is the standard deviation of the depth measurement along the ray at a given pixel coordinate, $\mathbf{\Sigma}_{\mathbf{p}}$ represents the pixel quantization error [12] and $\mathbf{J}_{\mathbf{x}}$ is the Jacobian matrix

$$\mathbf{J}_{\mathbf{x}} = \begin{bmatrix} f_x^{-1} & 0 & (u - c_x) f_x^{-1} \\ 0 & f_y^{-1} & (v - c_y) f_y^{-1} \\ 0 & 0 & 1 \end{bmatrix} \tag{9}$$

The value of σ_z is computed for each pixel in the sensor measurement from an empirically fitted model presented in [17] for a Kinect RGB-D sensor, where σ_z is shown to predominantly vary with the z coordinate of a point $\mathbf{x}_z$,

$$\sigma_z = 0.0012 + 0.0019 \left(\mathbf{x}_z - 0.4\right)^2 + \frac{0.0001}{\sqrt{\mathbf{x}_z}} \tag{10}$$

V. Results

In this section, we demonstrate the ability of our proposed map representation to be used for robotics applications such as dense incremental surface reconstruction, map compression and pose estimation. We evaluate the accuracy and performance of our proposed mapping strategy on multiple datasets quantitatively and qualitatively. First, we demonstrate the correctness of our incremental mapping strategy on noisy input data. Second, we compare the metric accuracy of our mapping approach to various state-of-the-art mapping algorithms with open source implementations and demonstrate the superior reconstruction performance with lower memory requirements. Third, we demonstrate the mapping strategy's compression capabilities while incurring marginal increase in reconstruction error. Fourth, we compare the qualitative performance of our proposed approach on publicly available real world datasets in terms of quality and error of reconstruction. Finally, we demonstrate the applicability of our map representation in solving robotics relevant problems such as frame-to-frame pose-estimation. Table I lists all the parameters used in all the following experiments, except in V-D.

We only compare the memory footprint utilized by NDTMap[1] and OctoMap[2] for storing occupied cells to maintain a fair comparison of compression capabilities. We use the following datasets for our evaluations:

1) D1: ICL-NUIM Living room Dataset [8, 3]
2) D2: Lounge Dataset, D3: Copyroom Dataset, D4: Stonewall Dataset [29]

[1] https://github.com/OrebroUniversity/perception_oru-release
[2] https://github.com/OctoMap/octomap

Parameters	b	α_λ cm^2	α_{len} cm^2	α_{conf}	α_n cm	N
Level-0	8	.11	2.8	0.1	1	40
Level-1	32	1.0	11.1			
Level-2	160	2.8	100			

TABLE I: Table of parameters

A. Comparison Metrics

The quantitative evaluation of our proposed 3D mapping approach is performed using the following metrics:

1) Reconstruction Error: Gaussian distributions are generative models. We can therefore sample 3D points from each Gaussian distribution in ${}^{w}\mathbf{\Theta}^{l}$ at hierarchical level-l, within 3σ confidence bounds to reconstruct a global 3D point cloud of the environment. The reconstruction error is then computed as the mean distance of each 3D point in the sampled point cloud to its closest surface in the ground truth mesh of the environment. For OctoMap, reconstruction error is defined as the distance of each occupied voxel center to the closest surface in the true mesh.
2) Precision of Reconstruction: Precision is defined as the fraction of sampled points whose distance to the ground truth mesh is less than $\alpha_{l,\lambda}$ for our proposed approach. For NDTMap and OctoMap precision is defined as the fraction of points whose distance to the ground truth mesh is less than the resolution of the map.
3) Recall of Reconstruction: Recall of the reconstruction is computed by sampling 3D points uniformly over the ground truth mesh and querying whether the points are observed by the 3D reconstructed map. For our proposed approach and NDTMap, a point is defined as observed if it lies inside the 3σ bounds of any Gaussian distribution in the map. For OctoMap, a point is defined as observed if it lies inside an observed voxel of the map.
4) Memory Consumption: The memory consumption of a representation is defined in MB as the required storage space for each map representation at the end of a dataset.

B. Accuracy of Representation

In this section, we demonstrate the accuracy of our incremental mapping strategy on perfect and noisy sensor data in D1 dataset. We compare the performance of Gaussian distributions fitted over individual sensor scans, to distributions fitted incrementally over perfect sensor data to distributions fitted incrementally over noisy sensor data. Table II shows that the proposed incremental strategy has low reconstruction error of the input information obtained from every scan. Further, actively incorporating the noise model of the sensor enables the framework to fit a more compressed representation over noisy data thus reducing the memory requirement of the map representation while increasing the precision of reconstruction over time.

Map Type	Error(m)	Precision	Recall	Size (MB)
Perfect input data				
Individual	0.0019	0.890	0.997	0.913
Incremental	**0.0006**	**0.987**	**0.996**	**0.0372**
Noisy input data				
Incremental	0.0031	0.790	0.985	0.2604
Noise Compensated	**0.0017**	**0.939**	**0.992**	**0.041**

TABLE II: Quantitative performance comparison of ${}^{w}\mathbf{\Theta}_t^0$ fit to on Dataset D1 with and without noise modeling. Incrementally updating the map reduces the reconstruction error and has lower storage complexity that fitting distributions over individual scan.

Approach	Error (m)	Precision	Recall	Size (MB)
Proposed	**0.0019**	**0.890**	**0.985**	**0.128**
NDTMap	0.0042	0.691	0.653	0.731
OctoMap	0.0236	0.251	0.636	0.5901

TABLE III: Quantitative comparison of state-of-the-art mapping approaches with the proposed approach on noisy dataset D1. OctoMap and NDTMap are being fit at 5 cm resolution and for the proposed approach $\alpha_{2,\lambda} = 5$ cm.

C. Surface Reconstruction Accuracy

We evaluate the surface reconstruction accuracy of our approach and compare it to state-of-the-art algorithms, OctoMap [9] and NDTMap [1], on the entire dataset D1 with added Gaussian noise to each sensor measurement according to a sensor model described in Sec. IV-B. We also compare the memory footprint of these representations and demonstrate the superior compression and representation capability of the proposed method. The fitted model represents the environment with high accuracy by representing input points as uncertain distributions. Table III shows that given current sensor pose estimates the proposed approach can represent the scene with higher precision, reconstruction accuracy and lower memory footprint than OctoMap and NDTMap.

D. Map Compression

In this section, we demonstrate the capability of the proposed mapping strategy to obtain a magnitude of compression of input data with minor loss of reconstruction accuracy. Table IV demonstrates that in structured environments, reducing the required surface reconstruction accuracy from 1 mm to 10 cm, reduces the memory requirement of the map by an order of magnitude while still achieving mean reconstruction error less than 1 cm.

E. Real World Datasets

The objective of using these datasets is to demonstrate the performance of our pipeline on real-world datasets where the provided ground truth poses are noisy and a correct sensor model is not available. D2, D3 and D4 datasets are captured using a hand-held sensor and sensor poses are estimated as described by Zhou and Koltun [29]. In order to compare the

$\alpha_{0,\lambda}$ (m)	$\alpha_{0,\text{len}}$ (m)	Error (m)	Precision	Recall	Size (MB)
0.01	0.05	0.0008	0.969	**0.989**	0.314
0.001	0.05	**0.0006**	**0.987**	**0.989**	0.349
0.1	0.1	0.0029	0.877	0.985	0.075
0.01	0.1	0.0016	0.934	0.985	0.195
0.001	0.1	0.0009	0.966	**0.989**	0.143
0.1	0.2	0.0098	0.701	0.970	**0.020**
0.01	0.2	0.0034	0.883	0.982	0.0308
0.001	0.2	0.0033	0.930	**0.989**	0.084

TABLE IV: Quantitative comparison of various performance metrics as different map resolutions. Increasing the thickness and length thresholds dramatically reduce the memory footprint of the proposed algorithm with small increase in reconstruction error.

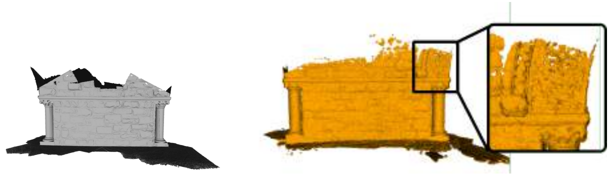

Fig. 5: Qualitative comparison of the model reconstructed by our approach at high resolutions on dataset D4: *Left:* Mesh provided by [29]; *Right:* proposed reconstruction with $\alpha_{0,\lambda} = 0.1$ cm. The zoomed in view shows that the Level-0 model is able to retain minute texture details in the scene and create a high quality scene reconstruction using succinct Gaussian distributions as surface primitives.

performance of OctoMap and NDTMap with the proposed hierarchical approach, we vary the voxel grid size for these map representations and compare the trade-off between memory consumption and reconstruction error of the representations at various configurations. However, since our approach does not use 3D voxels to change the resolution, we compare the performance of our approach at all three hierarchical levels with OctoMap and NDTMap at multiple resolutions. Table VI and Table V demonstrate that for D2 and D3 datasets, the proposed hierarchical map representation outperforms NDTMap and OctoMap in both compactness of the representation and the accuracy of reconstruction at all three hierarchical levels. Even with noisy pose estimates and unknown sensor noise model the proposed algorithm creates a qualitatively accurate representation of the environment and is orders of magnitude more memory efficient than NDTMap and OctoMap on these datasets. NDTMap fits volumetric distributions to fixed sized voxels, while OctoMap represents the world as voxels. Our approach on the other hand, can more accurately capture the spread of the structure data along planar regions and shown in Fig. 7. Therefore, we can achieve very high compression values while retaining high reconstruction accuracy. The proposed map refinement step eliminates the spurious distributions learnt on noisy sensor observations. Therefore, our approach as a low reconstruction error and a small standard deviation, whereas NDTMap has a large amount of noise incorporated in the final map representation as shown in Fig. 6. These figures show that our approach achieves the best trade-off between memory consumption and reconstruction accuracy and is capable of reconstructing the world with high accuracy. We also qualitatively compare a level-0 model reconstruction of D4 dataset at $\alpha_{0,\lambda} = 0.1$ cm in Fig. 5. The level-0 model retains minute details in the scene but requires a magnitude more memory than the level-2 model.

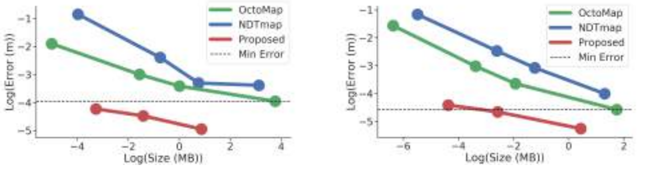

Fig. 6: Quantitative comparison of the memory vs accuracy trade-off of state-of-the-art mapping approaches with the proposed approach on D2 (left) and D3 (right) datasets. Our proposed approach has orders of magnitude lower reconstruction error than NDTMap and OctoMap at a fixed map storage complexity.

Approach	Resolution	Error (cm)	Size (MB)	Time(sec)
NDTMap	0.01	3.39 ± 8.01	22.8664	0.73 ± 0.08
	0.05	3.67 ± 20.67	2.1333	0.28 ± 0.02
	0.1	9.14 ± 37.09	0.4732	0.14 ± 0.01
	0.5	42.73 ± 97.83	0.0188	0.03
OctoMap	0.01	1.92 ± 1.85	43.3226	0.35 ± 0.12
	0.05	3.29 ± 2.83	1.0096	0.02
	0.1	4.99 ± 5.46	0.2104	0.01
	0.5	14.96 ± 20.59	**0.0066**	**0.01**
Proposed	Level 0	**0.07 ± 1.00**	2.3918	0.06 ± 0.01
	Level 1	**1.14 ± 1.39**	0.2433	
	Level 2	**1.46 ± 1.99**	0.0379	

TABLE V: Quantitative comparison of memory vs accuracy trade-off of state-of-the-art mapping approaches with the proposed approach on D2 dataset.

Approach	Resolution	Error (cm)	Size (MB)	Time (sec)
NDTMap	0.01	1.82 ± 8.35	3.69	0.74 ± 0.06
	0.05	4.6 ± 20.56	0.29	0.15 ± 0.01
	0.1	8.44 ± 30.01	0.07	0.08 ± 0.01
	0.5	31.16 ± 57.60	0.01	0.02
OctoMap	0.01	1.03 ± 1.97	5.71	0.05 ± 0.01
	0.05	2.61 ± 5.51	1.45	0.01
	0.1	4.87 ± 8.97	0.03	0.01
	0.5	20.86 ± 19.04	**0.002**	**0.01**
Proposed	Level 0	**0.52 ± 0.08**	1.55	0.05 ± 0.02
	Level 1	**0.94 ± 1.4**	0.07	
	Level 2	**1.21 ± 1.93**	0.01	

TABLE VI: Quantitative comparison of state-of-the-art mapping approaches with proposed approach on D3 dataset using reconstruction accuracy and memory footprint metrics.

F. Run Time Analysis

As seen in Fig 8, our mapping framework operates at an average frequency of 16 Hz over all the datasets. Correspondence computation is an essential step of our map update framework which is performed every time a new sensor measurement is observed. As the number of distributions in the scene increase, the time taken to render this index image as described in Sec. III-C1 also increases. However, as shown in Fig. 9 only a small subset of Level-0 Gaussian distributions are active.

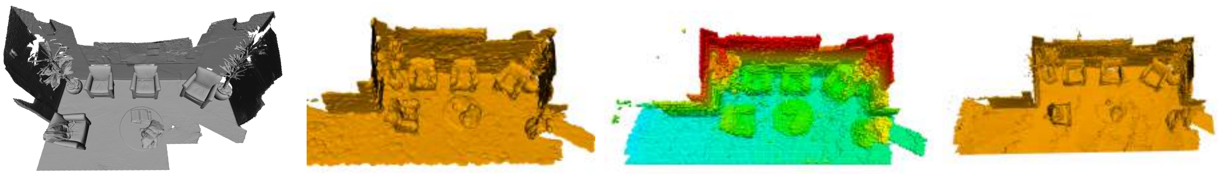

Fig. 7: Qualitative comparison of mapping approaches on Dataset D2. *Left to right:* Mesh estimated from [29]; proposed reconstruction framework; OctoMap, NDTMap. Unlike, OctoMap and NDTMap our map representation does not make assumptions about the distribution of the structure points and fits volumetric planar primitives to the data. Further, due to the explicit incorporation of sensor uncertainty in the representation, the proposed representation looks less cluttered and more structured than NDTMap and OctoMap while achieving orders of magnitude higher compression than either of those representations.

Using the active-inactive segmentation over Level-2 regions, we maintain a roughly constant number of active Level-0 distributions for correspondence computation and therefore, the proposed algorithm is real-time viable on a desktop grade CPU. The map updates can be further parallelized on a GPU due to the independent nature of our map representation, enabling the proposed algorithm to run at sensor rates (30 Hz for a COTS Kinect sensor).

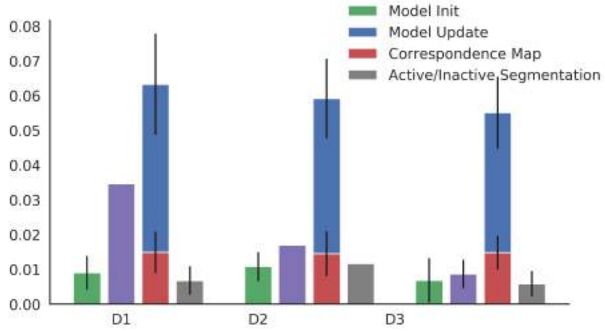

Fig. 8: Execution time comparison for the various subcomponents of the proposed Hierarchical GMM mapping algorithm on an i7 laptop grade CPU on datasets, D1, D2, and D3. Execution time for correspondence computation, that is a part of the model update step, scales with the number of Gaussian distributions. However, segmenting components as active/inactive keeps the execution time constant.

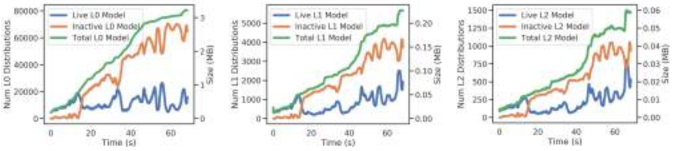

Fig. 9: Progression of the number of total, active and inactive Gaussian distributions at Level-0, Level-1 and Level-2 with time over dataset D1. Level-2 has orders of magnitude less distributions and therefore, can be efficiently used to sub-select Level-0 distributions that are within the sensor FOV.

G. Map Application: Pose Estimation

To demonstrate the usefulness of our map representation, we perform a simple frame-to-model registration. A point-to-closest planar distribution alignment process similar to the approach proposed by [22] is implemented. We use the incremental estimate of our level-1 hierarchical map, ${}^{w}\boldsymbol{\Theta}_t^1$ at each time step for normal computation, as it smoothens out the frame-to-model alignment cost function. Fig 10 shows a comparison between the estimated trajectory using our frame-to-model pose estimation with the ground truth trajectory. The global translation RMSE observed over a trajectory of length 37.15 m is 0.011 m and angular RMSE is $0.27°$ on the dataset D1.

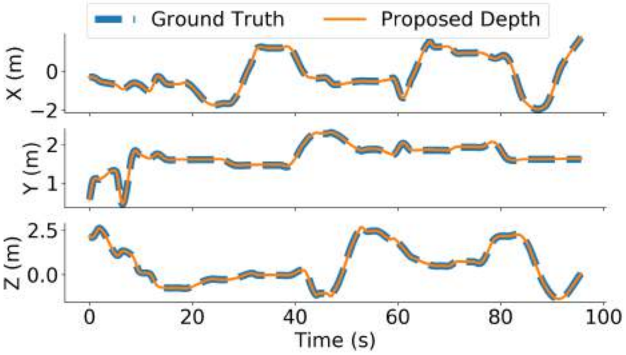

Fig. 10: Gaussian distributions provide a smooth cost function for accurate convergence of an ICP-like pose estimation algorithm. On dataset D1, a model-to-frame Point-to-Distribution tracking approach accumulates a global error of only 0.011 m over a trajectory of length 37.15 m.

VI. Discussion

We present an incremental, high fidelity and memory efficient surface mapping framework by exploiting prior understanding of the data acquisition process of a depth sensor. Our main contribution in this work is to show that by constraining the Gaussian distributions to model only surfaces, we are able to create a very accurate and succinct representation of the world. Further, by explicitly incorporating the noise characteristics of the sensor, we are able to create a map that best represents all the sensor observations with a low model complexity.

Future steps will involve extending this formulation to a real-time SLAM system by performing frame-to-model tracking and model-to-model pose refinements as hinted in Sec. V-G. We also intend to extend the formulation for very large scale mapping in real time on mobile robotic platforms.

References

[1] Peter Biber and Wolfgang Straßer. The normal distributions transform: A new approach to laser scan matching. In *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)(Cat. No. 03CH37453)*, volume 3, pages 2743–2748. IEEE, 2003.

[2] Paul Bromiley. Products and convolutions of Gaussian probability density functions. *Tina-Vision Memo*, 3(4):1, 2003.

[3] Sungjoon Choi, Qian-Yi Zhou, and Vladlen Koltun. Robust reconstruction of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5556–5565, 2015.

[4] Aditya Dhawale, Kumar Shaurya Shankar, and Nathan Michael. Fast Monte-Carlo Localization on Aerial Vehicles Using Approximate Continuous Belief Representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5851–5859, 2018.

[5] Aditya Dhawale, Xuning Yang, and Nathan Michael. Reactive Collision Avoidance Using Real-Time Local Gaussian Mixture Model Maps. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3545–3550. IEEE, 2018.

[6] Benjamin Eckart, Kihwan Kim, Alejandro Troccoli, Alonzo Kelly, and Jan Kautz. Accelerated generative models for 3D point cloud data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5497–5505, 2016.

[7] Alberto Elfes. Using occupancy grids for mobile robot perception and navigation. *Computer*, 22(6):46–57, 1989.

[8] Ankur Handa, Thomas Whelan, John McDonald, and Andrew J Davison. A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM. In *2014 IEEE international conference on Robotics and automation (ICRA)*, pages 1524–1531. IEEE, 2014.

[9] Armin Hornung, Kai M Wurm, Maren Bennewitz, Cyrill Stachniss, and Wolfram Burgard. OctoMap: An efficient probabilistic 3D mapping framework based on octrees. *Autonomous robots*, 34(3):189–206, 2013.

[10] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, 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*, pages 559–568, 2011.

[11] Bing Jian and Baba C Vemuri. Robust point set registration using gaussian mixture models. *IEEE transactions on pattern analysis and machine intelligence*, 33(8):1633–1645, 2010.

[12] Behzad Kamgar-Parsi and WA Sander. Quantization error in spatial sampling: comparison between square and hexagonal pixels. In *Computer Vision and Pattern Recognition, 1989. Proceedings CVPR'89., IEEE Computer Society Conference on*, pages 604–611. IEEE, 1989.

[13] Soo Min Kang and Richard P Wildes. The n-distribution Bhattacharyya coefficient. *EECS-2015-02*, 2015.

[14] Tobias Lang, Christian Plagemann, and Wolfram Burgard. Adaptive Non-Stationary Kernel Regression for Terrain Modeling. In *Robotics: Science and Systems*, volume 6, 2007.

[15] Bhoram Lee, Clark Zhang, Zonghao Huang, and Daniel D Lee. Online continuous mapping using Gaussian process implicit surfaces. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 6884–6890. IEEE, 2019.

[16] Martin Magnusson, Achim Lilienthal, and Tom Duckett. Scan registration for autonomous mining vehicles using 3D-NDT. *Journal of Field Robotics*, 24(10):803–827, 2007.

[17] Chuong V Nguyen, Shahram Izadi, and David Lovell. Modeling Kinect sensor noise for improved 3D reconstruction and tracking. In *2012 second international conference on 3D imaging, modeling, processing, visualization & transmission*, pages 524–530. IEEE, 2012.

[18] Matia Pizzoli, Christian Forster, and Davide Scaramuzza. Remode: Probabilistic, monocular dense reconstruction in real time. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2609–2616. IEEE, 2014.

[19] Jann Poppinga, Narunas Vaskevicius, Andreas Birk, and Kaustubh Pathak. Fast plane detection and polygonalization in noisy 3D range images. In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3378–3383. IEEE, 2008.

[20] Pedro F Proenca and Yang Gao. Probabilistic RGB-D odometry based on points, lines and planes under depth uncertainty. *Robotics and Autonomous Systems*, 104:25–39, 2018.

[21] Thomas Schops, Torsten Sattler, and Marc Pollefeys. BAD SLAM: Bundle adjusted direct RGB-D SLAM. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 134–144, 2019.

[22] Jacopo Serafin and Giorgio Grisetti. NICP: Dense normal based point cloud registration. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 742–749. IEEE, 2015.

[23] Alex Spitzer, Xuning Yang, John Yao, Aditya Dhawale, Kshitij Goel, Mosam Dabhi, Matt Collins, Curtis Boirum, and Nathan Michael. Fast and agile vision-based flight with teleoperation and collision avoidance on a multirotor. In *International Symposium on Experimental Robotics*, pages 524–535. Springer, 2018.

[24] Shobhit Srivastava and Nathan Michael. Approximate continuous belief distributions for precise autonomous inspection. In *2016 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, pages 74–80. IEEE, 2016.

[25] Wennie Tabib, Cormac OMeadhra, and Nathan Michael. On-manifold GMM registration. *IEEE Robotics and Automation Letters*, 3(4):3805–3812, 2018.

[26] Wennie Tabib, Kshitij Goel, John Yao, Mosam Dabhi, Curtis Boirum, and Nathan Michael. Real-Time Information-Theoretic Exploration with Gaussian Mixture Model Maps. *Proc. of Robot.: Sci. and Syst., FreiburgimBreisgau, Germany*, 2019.

[27] Sebastian Thrun, Christian Martin, Yufeng Liu, Dirk Hahnel, Rosemary Emery-Montemerlo, Deepayan Chakrabarti, and Wolfram Burgard. A real-time Expectation-Maximization algorithm for acquiring multiplanar maps of indoor environments with mobile robots. *IEEE Transactions on Robotics and Automation*, 20(3):433–443, 2004.

[28] Thomas Whelan, Stefan Leutenegger, R Salas-Moreno, Ben Glocker, and Andrew Davison. ElasticFusion: Dense SLAM without a pose graph. Robotics: Science and Systems, 2015.

[29] Qian-Yi Zhou and Vladlen Koltun. Dense scene reconstruction with points of interest. *ACM Transactions on Graphics (ToG)*, 32(4):1–8, 2013.