

# Modeling Uncertainty in 3D Gaussian Splatting through Continuous Semantic Splatting

Joey Wilson, Marcelino Almeida, Min Sun, Sachit Mahajan, Maani Ghaffari, Parker Ewen, Omid Ghasemalizadeh, Cheng-Hao Kuo, Arnie Sen

**Abstract**—In this paper, we present a novel algorithm for probabilistically updating and rasterizing semantic maps within 3D Gaussian Splatting (3D-GS). Although previous methods have introduced algorithms which learn to rasterize features in 3D-GS for enhanced scene understanding, 3D-GS can fail without warning which presents a challenge for safety-critical robotic applications. To address this gap, we propose a method which advances the literature of continuous semantic mapping from voxels to ellipsoids, combining the precise structure of 3D-GS with the ability to quantify uncertainty of probabilistic robotic maps. Given a set of images, our algorithm performs a probabilistic semantic update directly on the 3D ellipsoids to obtain an expectation and variance through the use of conjugate priors. We also propose a probabilistic rasterization which returns per-pixel segmentation predictions with quantifiable uncertainty. We compare our method with similar probabilistic voxel-based methods to verify our extension to 3D ellipsoids, and perform ablation studies on uncertainty quantification and temporal smoothing.

## I. INTRODUCTION

In order to plan, robots require a world model which captures geometric detail and higher levels of information about their environment. Although some papers propose mapless navigation [1]–[3], maps are still widely used due to an interpretable world model which temporally adapts as robots explore their surroundings. Depending on the robot application, maps can store different types of information to increase scene understanding.

For many robotic applications, uncertainty of the map is necessary to ensure safe planning in safety-critical environments. In these situations, robots must understand not only the type and location of objects, but confidence in the predictions as well. Uncertainty can arise from noisy perception networks, sensor noise, and sparse views which can ultimately result in incomplete maps.

Continuous mapping combats sparse data by leveraging spatial relations of points to fill in gaps in the map from sparse data probabilistically and with quantifiable uncertainty [4], [5]. Continuous mapping has been successfully applied to applications such as elevation mapping [6] and semantic mapping, by incorporating measurements into nearby cells in the robotic map through a kernel [7]. The kernel effectively

J. Wilson, M. Ghaffari, and P. Ewen are with the University of Michigan, Ann Arbor, MI 48109, USA. {wilsoniv, maanigj, pewen}@umich.edu

M. Ghaffari was supported by the DARPA TIAMAT project.

M. Almeida, S. Mahajan, M. Sun, O. Ghasemalizadeh, C. Kuo, and A. Sen are with Amazon Lab 126, Sunnyvale, CA, 94089, USA. {mmalmeid, msachit, minnsun}@amazon.com, {ghasemal, chkuo, senarnie}@amazon.com

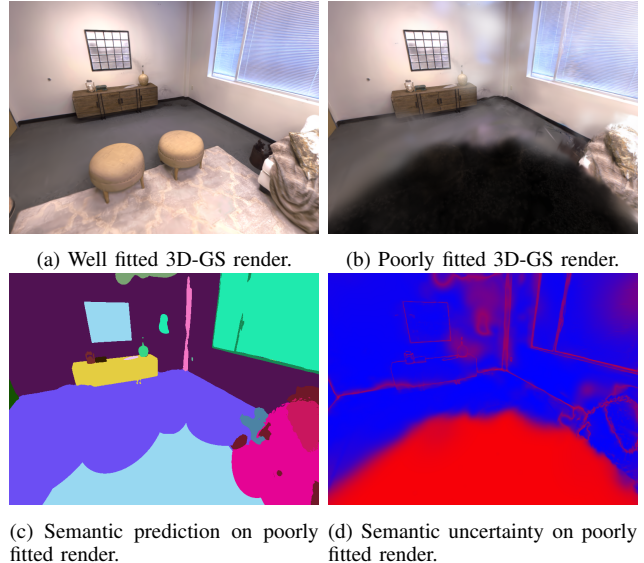


Fig. 1: While 3D-GS may provide high quality renderings of the environment at novel views with sufficient training data, it may fail to render views which are occluded, unseen, or at different angles from the training data. In the above image, CSS produces semantic (c) and RGB (b) predictions at a novel view without sufficient training data, resulting in a blurry render and incorrect segmentation. Through probabilistic inference, CSS identifies blurs and gaps in the render which correlate with reconstruction quality (d).

defines the influence of input points over nearby cells probabilistically, leading to a closed form update solution through Bayesian Kernel Inference (BKI). However, one challenge of BKI is defining the kernel function, which is generally hand-crafted and recently was shown to be learnable, resulting in 3D ellipsoid shapes [8]. Additionally, BKI has been limited to grid-based solutions which are prone to discretization errors and require accurate depth estimation.

Separately, 3D Gaussian Splatting (3D-GS) proposes a new method for novel view synthesis, which *learns* to model the world explicitly as 3D ellipsoids, with high quality renderings from any angle without the discretization error of grid-based map representations [9]. 3D-GS has captured the attention of the robotics community, with many methods proposing to add additional features to 3D-GS [10], [11] and incorporate 3D-GS into simultaneous localization and mapping (SLAM) pipelines [12], [13]. Since 3D-GS can fail to rasterize at novel views without warning, several recent works have examined uncertainty quantification 3D-GS. Compared to recent works leveraging Fisher Information [14], [15] or variational inference [16], our approach quantifies uncertainty through conjugate priors and categorical

dissonance.

In this work, we leverage the insight that 3D-GS learns valid kernels to propose a novel method for uncertainty quantification in 3D-GS. Our method, which we call Continuous Semantic Splatting (CSS), incorporates semantically labelled images in a Bayesian framework to capture the semantic uncertainty of each 3D ellipsoid. Additionally, through a novel rasterization method, we capture the semantic variance from noisy segmentation predictions in pixel space, as well as information on conflicting categories caused by poor renderings at novel views. To summarize, our contributions are:

- i. Extend continuous mapping literature from voxel grids to 3D-GS world representation.
- ii. Formulate novel 3D-GS semantic update with quantifiable semantic variance of ellipsoids.
- iii. Probabilistic semantic 3D-GS rasterization with quantifiable uncertainty.

## II. RELATED WORK

In this section we briefly review relevant literature on probabilistic semantic mapping and continuous Bayesian Kernel Inference (BKI), which approaches the ellipsoid world model representation of 3D-GS with quantifiable uncertainty however lacks expressive kernels and remains limited to voxel structures. Next, we present background on the recently developed 3D-GS world model representation, which *learns* an ellipsoid world model representation with rotation yet lacks innate uncertainty quantification.

### A. Probabilistic 3D Semantic Mapping

In 3D semantic mapping, the goal of the algorithm is to receive sequences of exteroceptive data and update a 3D model of the world with semantic labels and quantifiable uncertainty. A direct approach to solve this problem is to leverage off-the-shelf neural networks to semantically label 3D exteroceptive data, and probabilistically update the map cells occupied by the point through either a voting scheme or Bayesian update [17]–[21]. However, this form of update may lead to sparse maps due to sparse 3D data. To counteract the sparsity of 3D data, Gaussian Processes (GP’s) can incorporate input points to form a more complete map [22]–[24]. However, GP’s have a cubic computational complexity with respect to the number of input points, rendering their use impractical in scenes with high amounts of sensory data. As a result, BKI [4] is widely used as an approximation of GP’s to efficiently model the influence of sparse points on the map through kernel inference [5]. Semantic BKI [7] extended the BKI framework to semantic labels, and has been applied to applications such as off-road driving [25], [26] where uncertainty is critical.

However, one limitation of Semantic BKI is kernel selection, as kernels are generally hand-crafted. ConvBKI proposes to learn the shape of the kernel through a neural network, resulting in 3D ellipsoid shapes per semantic category [8]. While the categorical shapes learned by ConvBKI can enable better informed continuous mapping, the

kernels are limited due to the discretization of voxels, the lack of a rotation on the 3D ellipsoid distribution of input points, as well as requiring pre-training on a separate set of semantic segmentation labels. To alleviate these limitations, we propose to extend continuous semantic mapping to 3D-GS, which learns 3D ellipsoid distributions with rotations on more readily available camera data.

### B. 3D Gaussian Splatting

3D Gaussian Splatting (3D-GS) [9] is a new method for novel view synthesis which represents the world explicitly as 3D ellipsoids with color, as opposed to previous methods for novel view synthesis which represented the world implicitly such as Neural Radiance Fields (NeRF’s) [27]. Given a set of input images, 3D-GS optimizes the number, location, shape, and color of a set of 3D ellipsoids to best represent the training images. 3D-GS does not require pixel-wise depth predictions of the training images to predict colors, due to a depthwise rasterization process known as alpha compositing. After training, the 3D-GS model can be used to render images at any pose with high quality [28].

Due to the explicit 3D ellipsoid representation and high rasterization quality, 3D-GS has been expanded through works which improve the semantic scene understanding of the model [10], [11], and apply the representation to classic robotics problems such as SLAM [12], [13]. Most works focusing on improving the semantic scene understanding of 3D-GS incorporate features from off-the-shelf segmentation network by learning to render images with features, however 3D-GS can still fail without warning at novel views.

Recent works have studied uncertainty quantification in 3D-GS through Fisher Information [14], [15] and variational inference [16]. While both approaches have shown success, variational inference requires extensive additional training and Fisher Information is most directly suited for information quantification such as active perception. Additionally, neither approach incorporates semantic information to identify categorical conflict. Therefore, we propose to bridge the gap between uncertainty quantification in probabilistic robotic mapping and 3D-GS by leveraging the insight that the 3D ellipsoid shape learned by 3D-GS is a valid kernel which can be incorporated into the BKI framework.

## III. METHOD

In this section, we introduce our method Continuous Semantic Splatting (CSS), which probabilistically updates and rasterizes 3D semantic predictions in the 3D-GS representation using continuous BKI. First, we introduce preliminaries on the 3D-GS ellipsoid representation. Next, we introduce preliminaries on BKI and demonstrate how our method extends BKI to 3D-GS. Finally, we present a method for rasterization of the 3D conjugate prior distributions which maintains quantifiable uncertainty in the pixel space.

### A. Preliminaries: Gaussian Splatting

3D Gaussian Splatting represents a scene through 3D ellipsoids with location  $\mu$ , opacity  $\alpha$ , rotation  $R$ , scale  $S$ ,

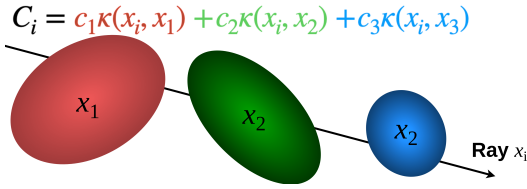


Fig. 2: 3D-GS renders pixels as a linear combination of 3D ellipsoids, where the influence of each 3D ellipsoid is determined by the spatial position and shape of the ellipsoid  $x_n$  relative to pixel  $x_i$  as  $\kappa(x_i, x_n)$ . We propose to leverage the learned expressive kernels of 3D-GS to perform a probabilistic semantic update and rasterization which enables uncertainty quantification.

and color  $c$ . Together, the scale and rotation define the shape of the ellipsoid,  $\Sigma$ , which determine the influence of the ellipsoid over pixels together with the ellipsoid’s location and opacity. To render a 2D image, 3D ellipsoids are first converted to a 2D ellipsoid with shape  $\Sigma'_n$  and location  $\mu'_n$  in a process known as splatting [9], [28]. Given a 2D pixel with location  $x'_i$  and a 2D splat  $n$ , the spatial contribution of the splat is first calculated through a kernel as:

$$k(x'_i, x'_n) = \exp\left(-\frac{1}{2}(x'_i - \mu'_n)^T \Sigma'^{-1}_n (x'_i - \mu'_n)\right), \quad (1)$$

where a pixel located at the center of the ellipsoid would result in a contribution of 1. This kernel can be combined with the ellipsoid’s opacity to obtain a measure of influence on a passing pixel:

$$\alpha'_n = \alpha_n \cdot k(x'_i, x'_n). \quad (2)$$

Depths of the 3D ellipsoids are incorporated into the 2D rendering through alpha compositing, which weights the contribution of each ellipsoid to the pixel’s color through depthwise sorting as:

$$\kappa(x'_i, x'_n) = \alpha'_n \prod_{j=1}^{n-1} (1 - \alpha'_j), \quad (3)$$

resulting in a final blended pixel color of:

$$C_i = \sum_{n=1}^N c_n \kappa(x'_i, x'_n). \quad (4)$$

Altogether,  $\kappa(x'_i, x'_n)$  defines the contribution of ellipsoid  $n$  to the total color  $C$ , as a function of the ellipsoid’s depth, shape, and opacity. Note that the contribution of the ellipsoid evaluates to a number between 0 and 1 in all cases, with a value of 1 when the pixel ray terminates exactly at the center of the 3D ellipsoid. Based on this insight, we propose that  $\kappa(x', \mu'_n)$  is a valid kernel that satisfies the constraints of BKI which we present next.

### B. Probabilistic Semantic Update

Given a fully trained 3D-GS model on a set of images  $\mathcal{I}$ , we propose to leverage the learned kernels of 3D-GS to perform a probabilistic Bayesian update on the semantic belief of each 3D ellipsoid with BKI. Compared with previous work, our method does not require additional training to learn to render features. Instead, CSS extends

classical probabilistic robotic mapping methods to 3D-GS with quantifiable uncertainty.

For each image in our training set, we first use an off-the-shelf neural network to obtain semantic segmentation predictions  $y_i$  for each pixel  $x_i$ , where  $y_i$  is a one-hot encoded vector. Given the training data, our goal is to learn the category of each ellipsoid by defining a categorical likelihood:

$$p(y_i|\theta_i) = \prod_{c=1}^C (\theta_i^c)^{y_i^c}. \quad (5)$$

However, in 3D-GS each pixel is rasterized through partial observations of many ellipsoids. Likewise, the semantic category of each pixel should influence the semantic belief of the same set of ellipsoids. Based on this insight, we propose to perform the Bayesian update with BKI, which relates the likelihood  $p(y_i|\theta_i)$  to the extended likelihood  $p(y_i|\theta_n, x_i, x_n)$  through a kernel as:

$$p(y_i|\theta_n, x_i, x_n) \propto p(y_i|\theta_i)^{\kappa(x_i, x_n)}. \quad (6)$$

The only requirements on the kernel function are that

$$0 \leq \kappa(x_i, x_n) \leq 1 \quad \text{and} \quad \kappa(x_i, x_n) = 1 \forall x_i = x_n, \quad (7)$$

which the 3D-GS kernel satisfies. Based on this observation, BKI provides a method to relate semantic segmentation of pixels to the semantic state of the ellipsoid the pixel passes through. Incorporating the categorical likelihood of a pixel into the extended likelihood formulation yields:

$$p(y_i|\theta_n, x_i, x_n) \propto \left[ \prod_{c=1}^C (\theta_i^c)^{y_i^c} \right]^{\kappa(x_i, x_n)}, \quad (8)$$

which effectively defines the semantic likelihood of a pixel according to the influence of ellipsoid  $n$  over the pixel. Next, we define a prior distribution over the semantic state of ellipsoid  $n$  using the conjugate prior of the categorical distribution, the Dirichlet distribution. The Dirichlet distribution defines a distribution over a distribution through concentration parameters  $\alpha$ , as:

$$p(\theta_n) \propto \prod_{i=1}^C \theta_{n,c}^{\alpha_n^c - 1}. \quad (9)$$

The concentration parameters model the counts of observations of each category, and can be decoded into an expected categorical distribution, and variance. Intuitively, more observations results in lower variance or uncertainty, and the probability of each category can be identified through normalization:

$$\mathbb{E}[\theta_n^c] = \frac{\alpha_n^c}{\sum_{j=1}^C \alpha_n^j}, \quad \text{Var}[\theta_n^c] = \frac{\mathbb{E}[\theta_n^c](1 - \mathbb{E}[\theta_n^c])}{1 + \sum_{j=1}^C \alpha_n^j}. \quad (10)$$

Combining the conjugate prior distribution and extended likelihood, the Bayesian update over the semantic category of ellipsoid  $n$  given all training pixels  $\mathcal{D}$  can be written as:

$$p(\theta_n | x_n, \mathcal{D}) \propto \left[ \prod_{i=1}^N \left[ \prod_{c=1}^C (\theta_n^c)^{y_i^c} \right]^{\kappa(x_n, x_i)} \right] \prod_{c=1}^C \theta_{n,c}^{\alpha_n^c - 1}, \quad (11)$$

which can be simplified to an un-normalized update of the concentrations parameters:

$$\alpha_n^c \leftarrow \alpha_n^c + \sum_{i=1}^N \kappa(x_i, x_n) y_i^c. \quad (12)$$

To summarize, given a set of images  $\mathcal{I}$  and a pre-trained 3D-GS model on the set of images, we first label each pixel  $x_i$  in the set of training images with an off-the-shelf semantic segmentation network to obtain per-pixel one-hot encoded predictions  $y_i$ . Next, we adopt an uninformative conjugate prior over the semantic category of all 3D ellipsoids [7]. Last, we update the concentration parameters of each 3D ellipsoid by computing an un-normalized sum over all pixels the ellipsoid influences with Eq. (12).

### C. Probabilistic Semantic Rasterization

While the BKI update is able to capture variance in the semantic segmentation input network, it does not capture uncertainty in novel views directly. We propose to leverage the rasterization process of 3D-GS to render semantic predictions with uncertainty in the pixel space. Our intuition is that 3D-GS rasterizations can fail in three ways which can be captured through semantic uncertainty. First, semantic rasterization can fail when the input segmentation network is noisy, resulting in high variance predictions. Second, at novel views there may be an absence of ellipsoids, in which case the extended likelihood would have high variance. Finally, novel views may have a blur of objects resulting in a high probability of conflicting categories. Therefore, we propose to model the categorical distribution of the rendered pixel as a linear combination of 3D ellipsoids:

$$\theta_i = \sum_{n=1}^N \kappa(x_i, x_n) \theta_n. \quad (13)$$

The expectation of the categorical variable can then be rasterized directly through the 3D-GS rasterization process:

$$\mathbb{E}(\theta_i) = \sum_{n=1}^N \kappa(x_i, x_n) \mathbb{E}(\theta_n), \quad (14)$$

and the variance can be similarly modeled under an assumption of independence between ellipses:

$$\text{Var}(\theta_i) = \sum_{n=1}^N \kappa(x_i, x_n)^2 \text{Var}(\theta_n). \quad (15)$$

When performing alpha compositing to render images, 3D-GS incorporates a background class to fill the absence of ellipsoids. In this case, the background is treated as another ellipsoid, however the weight of the background color is:

$$\kappa(x_i, x_b) = 1 - \sum_{n=1}^N \kappa(x_i, x_n), \quad (16)$$

where  $x_b$  is the background. Similarly, we propose to incorporate a background distribution, where  $\theta_b \sim \text{Dir}(\alpha_b)$  and  $\alpha_b$  is a small positive number uniformly distributed for each category, such that the background has high semantic uncertainty. Due to the formulation of our problem within 3D-GS, variance and expectation both provide important measures of uncertainty with a trade-off in information about conflicting ellipsoid categories (expectation), or few observations (variance).

### D. Uncertainty at Image Level

From the pixels, we may also desire uncertainty at the image level for tasks such as active perception. Inspired by D-Optimality [29], [30], which defines a functional of the covariance matrix to estimate information, we compute the uncertainty of an image  $\mathcal{I}$  from the variance as:

$$U(\mathcal{I}_{\text{Var}}) = \sqrt[n]{|\Sigma_{\mathcal{I}}|} \\ = \exp\left(\frac{1}{n} \sum_{i=1}^n \log(\text{Var}(\hat{\theta}_i))\right), \quad (17)$$

where  $n$  is the number of pixels in the image and  $\hat{\theta}$  is the categorical variable indexed by the most likely category. We can also obtain a measure of uncertainty from the expectation as:

$$U(\mathcal{I}_{\mathbb{E}}) = 1 - \frac{\sum_{i=1}^n \mathbb{E}(\hat{\theta}_i)}{n}, \quad (18)$$

where the sign is flipped to obtain the uncertainty. This heuristic for uncertainty indicates the pixel-wise average of the probability mass for all non-predicted categories. Intuitively, a low probability mass of non-predicted categories for a pixel indicates low confidence in the predicted categories.

## IV. RESULTS

In this section, we verify the probabilistic update and rasterization of our method by comparing to similar voxel-based approaches. We demonstrate that our application of kernel inference to 3D-GS is valid experimentally, as demonstrated by comparable precision, with the benefit of a more complete representation without a requirement of accurate depth. Next, we study the uncertainty quantification capabilities of our method on semantic variance caused by noisy segmentation networks, as well as image rasterization errors caused by insufficient training data. Finally, we study the smoothing effect of our model and the performance gap of our method with perfect segmentation.

### A. Comparison to Voxel-Based Methods

Following the experimental setup of Semantic BKI (S-BKI) [7] which our work is built on, we compare our approach to several probabilistic voxel-based approaches on the KITTI driving dataset [31] to validate our approach against similar mapping algorithms. All methods are provided the same set of images and corresponding semantic segmentation predictions [32]. Whereas the voxel-based approaches require pixel-wise depth predictions estimated from ELAS [33], our approach requires pre-training to learn the

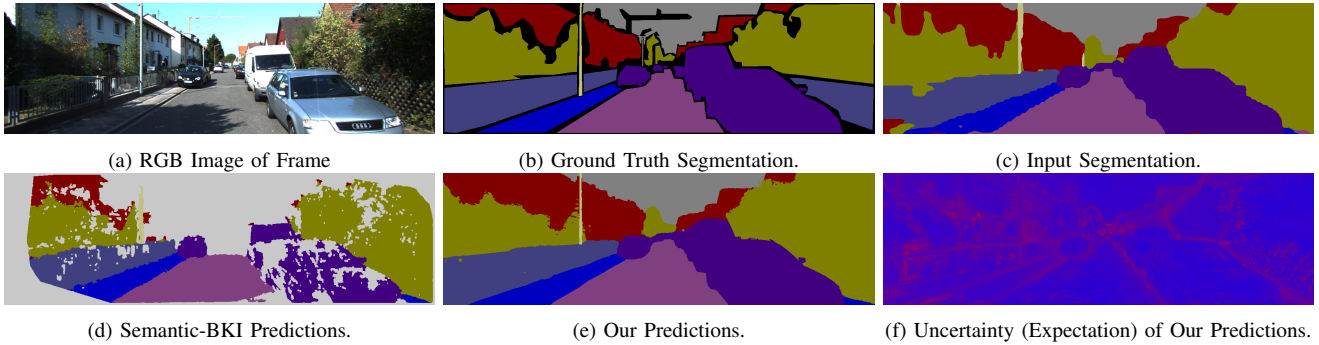


Fig. 3: Comparison of our method to a probabilistic voxel baseline on the KITTI driving dataset. Our method achieves similar segmentation results on the pixels predicted by the voxel method, and predicts more of the scene due to the lack of a requirement for accurate depth. Additionally, our method does not have discretization, which is beneficial for fine categories such as poles.

structure of the scene. Models are evaluated on the mean Intersection over Union (mIoU) metric of per-pixel semantic segmentation predictions. The most direct comparisons are Semantic BKI [7], which applies the BKI operation on voxels with a spherical kernel, and ConvBKI [8] which applies the BKI update on voxels with learned per-category kernels.

First, we compare our approach to voxel-based approaches on all pixels within each image. Qualitative results are shown in Fig. 3, and quantitative results are shown in Table I. Whereas voxel-based approaches are unable to complete the entire scene due to requiring accurate per-pixel depth, 3D-GS is capable of rendering the entire scene without depth. This difference is visible qualitatively by comparing the gaps in the voxel map generated by Semantic BKI in Fig. 3 to the semantic rendering produced by our method which does not have any gaps. The quantitative evaluation in Table I also supports this claim, as our method achieves a higher mIoU than all probabilistic baselines, as well as improving upon the input segmentation network, highlighting the ability of our method to incorporate sequences of images. While Table I demonstrates improvement in completion, we would like to note that the performance of the voxel-based methods is correlated with the accuracy of depth predictions, and may improve with better depth estimation algorithms.

Therefore, to understand the ability of our method to incorporate semantic measurements into a 3D model more directly compared to voxel-based methods, we quantitatively compare our model over the same set of masked pixels produced by the probabilistic approaches. Since all algorithms are provided the same set of data, we would expect the results on semantic segmentation produced by the map to be similar. This is visually apparent in Fig. 3, as well as demonstrated

TABLE I: Results on KITTI Odometry sequence 15 [31]. 3D-GS is able to complete more of the scene since it does not rely on accurate depth estimation for training or rasterization.

Method	Building	Road	Vege.	Sidewalk	Car	Sign	Fence	Pole	mIoU (%)
Segmentation [32]	92.1	93.9	<b>90.7</b>	81.9	94.6	19.8	78.9	<b>49.3</b>	75.1
Yang et al. [34]	32.9	85.8	59	79.3	61	0.9	46.8	33.9	50
BGKOctoMap-CRF [5]	50	86.6	64.1	74.9	61	0.0	47.5	36.7	52.6
S-CSM [7]	42.6	87.3	62.9	77.9	62.6	17.1	47.7	34.8	54.1
S-BKI [7]	49.3	88.8	69.1	78.2	63.6	22	49.3	36.7	57.1
Ours	<b>95.5</b>	<b>95.8</b>	89.2	<b>84.8</b>	<b>95.5</b>	<b>25.4</b>	<b>80.8</b>	45.1	<b>76.5</b>

TABLE II: Results on KITTI Odometry sequence 15 [31] of masked pixels which have predictions from S-BKI.

Method	Building	Road	Vege.	Sidewalk	Car	Sign	Fence	Pole	mIoU (%)
Segmentation [32]	92.1	93.9	90.7	81.9	94.6	19.8	78.9	49.3	75.1
Yang et al. [34]	95.6	90.4	<b>92.8</b>	70.0	94.4	0.1	<b>84.5</b>	49.5	72.2
BGKOctoMap-CRF [5]	94.7	93.8	90.2	81.1	92.9	0.0	78.0	49.7	72.5
S-CSM [7]	94.4	95.4	90.7	84.5	95.0	22.2	79.3	51.6	76.6
S-BKI [7]	94.6	95.4	90.4	84.2	95.1	<b>27.1</b>	79.3	51.3	77.2
ConvBKI [8]	94.0	<b>95.6</b>	91.0	<b>87.2</b>	95.1	22.8	81.9	54.3	<b>77.7</b>
Ours	<b>95.6</b>	94.9	90.7	84.8	<b>95.3</b>	8.8	79.8	<b>60.8</b>	76.3

quantitatively in Table II, where results are mixed between categories. These results confirm our intuition that CSS performs a valid probabilistic update and rasterization.

One significant difference between our method and voxel-based BKI methods is in the sign and pole categories, where our method achieves a higher mIoU on the pole category but a lower mIoU on the sign category than ConvBKI and Semantic BKI. From examining the confusion matrices of both methods, we find that our method is prone to mislabeling the sign as a pole. This likely occurs because 3D-GS learns to combine visually similar categories into the same structure, whereas the ground truth of the KITTI dataset separates the sign and sign-post as two separate categories. Combining the sign and pole into one pole category, shown in Table III, we find that our method outperforms Semantic BKI on mIoU, and particularly on the combined pole and sign category. This result demonstrates that 3D-GS is more suitable for complex and detailed environments, where discretization from voxels cannot adequately represent thin objects.

### B. Uncertainty Quantification

We perform studies on an indoor environment created with the Replica simulator, which offers high-quality images and ground truth semantic labels [35], [36] suitable for testing

TABLE III: Results on KITTI Odometry sequence 15 [31] of masked pixels which have predictions from S-BKI. Compared with Table II, we combine the pole and sign categories into a single pole category.

Method	Building	Road	Vege.	Sidewalk	Car	Pole/Sign	Fence	mIoU (%)
S-BKI [7]	94.5	<b>96.2</b>	90.2	<b>87.5</b>	95.2	50.7	<b>80.1</b>	84.9
Ours	<b>95.6</b>	94.9	<b>90.7</b>	84.8	<b>95.3</b>	<b>61.4</b>	79.8	<b>86.1</b>

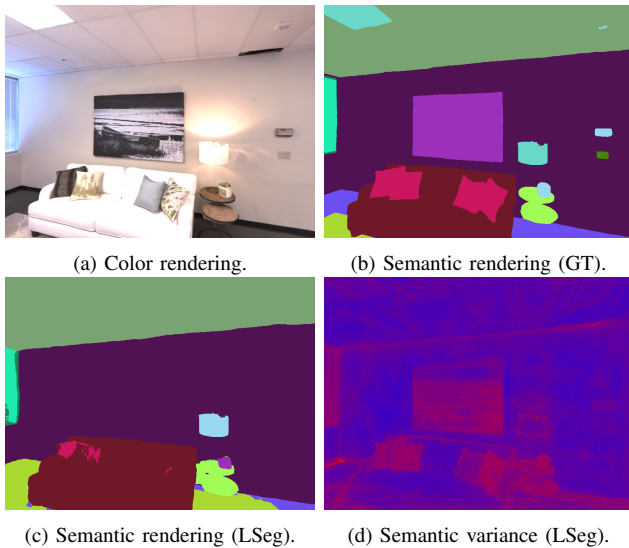


Fig. 4: Rasterizations from our method on an indoor environment. Our method achieves high quality semantic renderings using ground truth segmentation, shown in (b). Even with noisy segmentation input our method is capable of improving the segmentation with temporal smoothing (c), and can quantify uncertainty (d).

3D-GS. In this experiment, 405 images are collected along a manually controlled trajectory in the *room\_0* scene in order to simulate a robot path. The data contains 28 unique classes, including basic classes such as wall and floor, as well as more difficult classes such as electrical outlets. An example of the data can be found in Fig. 4. In order to study uncertainty quantification, we group predictions into bins by confidence, and iteratively remove the least confident sets of predictions, known as sparsification [37]. If uncertainty is properly calibrated, we would expect to see the performance improve as less confident predictions are removed.

**Uncertainty of Segmentation** To analyze the ability of our method to quantify uncertainty in the segmentation rasterizations caused by noisy input segmentation networks, we generate segmentation predictions for every image with the open-dictionary segmentation model LSeg [38]. Next, we train a 3D-GS model on all images of the dataset and update the concentration parameters of the 3D-GS model with the LSeg predictions. From the final concentration parameters, we generate pixel-level segmentation predictions for every image, as well as uncertainty calculated by the expectation and variance. Plotting the sparsification in Fig. 5, we find that both variance and expectation are highly correlated with

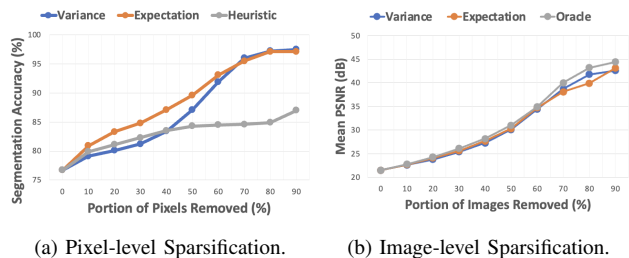


Fig. 5: Sparsification plot of pixel-level and image-level uncertainty. Uncertainty quantification from the expectation and variance are effective at both the image and pixel level.

Method	mIoU (%)	Accuracy (%)
LSeg	25.5	70.7
Ours (LSeg Segmentation)	<b>28.4</b>	<b>76.7</b>
Ours (GT Segmentation)	73.7	97.0

segmentation accuracy. Additionally, we compare against a heuristic baseline which computes confidence proportional to the amount of times ellipsoids have been observed. Concretely, pixel-wise confidence is computed as a weighted average of the Dirichlet normalization constants of contributing ellipsoids.

**Uncertainty of View** Next, we repeat the experiment at the image level, comparing image uncertainty from expectation and variance with the PSNR of the images using the predictions from LSeg. Instead of providing all images from the sequence to train the 3D-GS model and update the concentration parameters, only the first 100 frames are provided as training data, so that we can examine uncertainty quantification on unseen views. For this experiment, we also compare against an oracle baseline, which quantifies uncertainty according to the actual PSNR of each image. Both methods of quantifying uncertainty show correlation with image-level PSNR near the performance of the oracle.

### C. Smoothing Effect

Last, we study the smoothing effect of the continuous mapping operation on the predictions from LSeg in Table IV. First, we compare the accuracy and mIoU of the predictions from our model with the predictions from LSeg. We find that our method improves over the predictions of LSeg in both metrics, due to temporal incorporation of the segmentation predictions. Next, we repeat the process with ground truth segmentation to examine the loss from using 3D-GS as a map representation. We find that the accuracy is close to 100%, however the mIoU suffers in a couple of categories. Similar to before, we find that visually similar classes are blended, such as the electrical outlet being labeled incorrectly as wall.

## V. CONCLUSION

In this paper, we introduced a novel method for probabilistically incorporating semantic segmentation predictions into a 3D-GS world model. Our method combines the uncertainty quantification abilities of classical robotic mapping methods with the modern ellipsoid representation of 3D-GS by noting that the ellipsoids learned by 3D-GS are a natural extension of continuous mapping. We also proposed novel methods for uncertainty quantification in 3D-GS at the pixel and image level, and found that both the expectation and variance can be useful metrics of uncertainty. For future work, we believe that integrating this method within online 3D-GS frameworks would be a valuable extension for robotic mapping. Additionally, while our proposed method operates on a discrete set of categories, the probabilistic update and rasterization may be extended to open-dictionary continuous splatting through a continuous conjugate prior. Finally, we would like to expand quantitative analysis to include more data, specifically for image-level sparsification to identify failure cases.

## REFERENCES

- [1] S. Casas, A. Sadat, and R. Urtasun, "MP3: A Unified Model to Map, Perceive, Predict and Plan," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 14 403–14 412.
- [2] F. Codevilla, M. Müller, A. López, V. Koltun, and A. Dosovitskiy, "End-to-End Driving Via Conditional Imitation Learning," in *Proc. IEEE Int. Conf. Robot. and Automation*, 2018, pp. 4693–4700.
- [3] H. L. Chiang, A. Faust, M. Fiser, and A. Francis, "Learning Navigation Behaviors End-to-End With AutoRL," *IEEE Robot. Autom. Letter.*, vol. 4, no. 2, pp. 2007–2014, 2019.
- [4] W. R. Vega-Brown, M. Doniec, and N. G. Roy, "Nonparametric Bayesian inference on multivariate exponential families," in *Proc. Advances Neural Inform. Process. Syst. Conf.*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27, 2014.
- [5] K. Doherty, T. Shan, J. Wang, and B. Englot, "Learning-Aided 3-D Occupancy Mapping with Bayesian Generalized Kernel Inference," *IEEE Trans. Robot.*, vol. 35, no. 4, pp. 953–966, 2019.
- [6] 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 Robot. Autom. Letter.*, vol. 7, no. 3, pp. 7083–7090, 2022.
- [7] L. Gan, R. Zhang, J. W. Grizzle, R. M. Eustice, and M. Ghaffari, "Bayesian Spatial Kernel Smoothing for Scalable Dense Semantic Mapping," *IEEE Robot. Autom. Letter.*, vol. 5, no. 2, pp. 790–797, 2020.
- [8] J. Wilson, Y. Fu, A. Zhang, J. Song, A. Capodici, P. Jayakumar, K. Barton, and M. Ghaffari, "Convolutional Bayesian Kernel Inference for 3D Semantic Mapping," in *Proc. IEEE Int. Conf. Robot. and Automation*, 2023, pp. 8364–8370.
- [9] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3D Gaussian Splatting for Real-Time Radiance Field Rendering," *IEEE Trans. Graph.*, vol. 42, no. 4, July 2023.
- [10] M. Qin, W. Li, J. Zhou, H. Wang, and H. Pfister, "LangSplat: 3D Language Gaussian Splatting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, June 2024, pp. 20 051–20 060.
- [11] S. Zhou, H. Chang, S. Jiang, Z. Fan, Z. Zhu, D. Xu, P. Chari, S. You, Z. Wang, and A. Kadambi, "Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 21 676–21 685.
- [12] H. Matsuki, R. Murai, P. H. Kelly, and A. J. Davison, "Gaussian Splatting SLAM," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, June 2024, pp. 18 039–18 048.
- [13] L. Zhu, Y. Li, E. Sandström, S. Huang, K. Schindler, and I. Armeni, "LoopSplat: Loop Closure by Registering 3D Gaussian Splats," *arXiv*, vol. abs/2408.10154, 2024.
- [14] W. Jiang, B. Lei, and K. Daniilidis, "FisherRF: Active View Selection and Uncertainty Quantification for Radiance Fields using Fisher Information," *arXiv*, vol. abs/2311.17874, 2023.
- [15] A. Hanson, A. Tu, V. Singla, M. Jayawardhana, M. Zwicker, and T. Goldstein, "PUP 3D-GS: Principled Uncertainty Pruning for 3D Gaussian Splatting," *arXiv*, vol. abs/2406.10219, 2024.
- [16] L. Savant, D. Valsesia, and E. Magli, "Modeling uncertainty for Gaussian Splatting," *arXiv*, vol. abs/2403.18476, 2024.
- [17] J. Stückler, N. Biresev, and S. Behnke, "Semantic mapping using object-class segmentation of RGB-D images," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst.*, 2012, pp. 3005–3010.
- [18] H. He and B. Upcroft, "Nonparametric semantic segmentation for 3D street scenes," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst.*, 2013, pp. 3697–3703.
- [19] J. McCormac, A. Handa, A. Davison, and S. Leutenegger, "SemanticFusion: Dense 3D semantic mapping with convolutional neural networks," in *Proc. IEEE Int. Conf. Robot. and Automation*, 2017, pp. 4628–4635.
- [20] S. Sengupta, E. Greveson, A. Shahrokni, and P. H. S. Torr, "Urban 3D semantic modelling using stereo vision," in *Proc. IEEE Int. Conf. Robot. and Automation*, 2013, pp. 580–585.
- [21] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*. MIT press, 2005.
- [22] J. Wang and B. Englot, "Fast, accurate gaussian process occupancy maps via test-data octrees and nested Bayesian fusion," in *Proc. IEEE Int. Conf. Robot. and Automation*, 2016, pp. 1003–1010.
- [23] S. T. O'Callaghan and F. T. Ramos, "Gaussian process occupancy maps," *Int. J. Robot. Res.*, vol. 31, no. 1, pp. 42–62, 2012.
- [24] M. G. Jadidi, L. Gan, S. A. Parkison, J. Li, and R. M. Eustice, "Gaussian Processes Semantic Map Representation," *arXiv*, vol. abs/1707.01532, 2017.
- [25] J. Wilson, Y. Fu, J. Friesen, P. Ewen, A. Capodici, P. Jayakumar, K. Barton, and M. Ghaffari, "ConvBKI: Real-Time Probabilistic Semantic Mapping Network with Quantifiable Uncertainty," *IEEE Trans. Robot.*, pp. 1–20, 2024.
- [26] J. Kim, J. Seo, and J. Min, "Evidential Semantic Mapping in Off-road Environments with Uncertainty-aware Bayesian Kernel Inference," 2024.
- [27] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis," in *Proc. European Conf. Comput. Vis.*, 2020, pp. 405–421.
- [28] B. Fei, J. Xu, R. Zhang, Q. Zhou, W. Yang, and Y. He, "3D Gaussian Splatting as New Era: A Survey," *IEEE Trans. Graph.*, pp. 1–20, 2024.
- [29] J. Kiefer, "General equivalence theory for optimum designs (approximate theory)," *The annals of Statistics*, pp. 849–879, 1974.
- [30] J. A. Placed, J. Strader, H. Carrillo, N. Atanasov, V. Indelman, L. Carlone, and J. A. Castellanos, "A Survey on Active Simultaneous Localization and Mapping: State of the Art and New Frontiers," *IEEE Trans. Robot.*, vol. 39, pp. 1686–1705, 2022.
- [31] S. Sengupta, E. Greveson, A. Shahrokni, and P. H. S. Torr, "Urban 3D semantic modelling using stereo vision," in *Proc. IEEE Int. Conf. Robot. and Automation*, 2013, pp. 580–585.
- [32] F. Yu and V. Koltun, "Multi-Scale Context Aggregation by Dilated Convolutions," in *Proc. Int. Conf. Learning Representations*, 2016.
- [33] A. Geiger, M. Roser, and R. Urtasun, "Efficient large-scale stereo matching," in *Proc. Asian Conf. Comput. Vis.*, 2011, pp. 25–38.
- [34] S. Yang, Y. Huang, and S. Scherer, "Semantic 3D occupancy mapping through efficient high order CRFs," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst.*, 09 2017, pp. 590–597.
- [35] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma, A. Clarkson, M. Yan, B. Budge, Y. Yan, X. Pan, J. Yon, Y. Zou, K. Leon, N. Carter, J. Briales, T. Gillingham, E. Mueggler, L. Pesqueira, M. Savva, D. Batra, H. M. Strasdat, R. D. Nardi, M. Goesele, S. Lovegrove, and R. Newcombe, "The Replica Dataset: A Digital Replica of Indoor Spaces," *arXiv*, 2019.
- [36] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik, D. Parikh, and D. Batra, "Habitat: A Platform for Embodied AI Research," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019.
- [37] E. Ilg, Ö. Çiçek, S. Galesso, A. Klein, O. Makansi, F. Hutter, and T. Brox, "Uncertainty Estimates and Multi-hypotheses Networks for Optical Flow," in *Proc. European Conf. Comput. Vis.*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., 2018, pp. 677–693.
- [38] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl, "Language-driven Semantic Segmentation," in *Proc. Int. Conf. Learning Representations*, 2022.