2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

On the Overconfidence Problem in Semantic 3D Mapping

Joao Marcos Correia Marques[1], *IEEE Student Member*, Albert J. Zhai[1],
Shenlong Wang[1], *IEEE Member*, and Kris Hauser[1], *IEEE Senior Member*

***Abstract*— Semantic 3D mapping, the process of fusing depth and image segmentation information between multiple views to build 3D maps annotated with object classes in real-time, is a recent topic of interest. This paper highlights the fusion overconfidence problem, in which conventional mapping methods assign high confidence to the entire map even when they are incorrect, leading to miscalibrated outputs. Several methods to improve uncertainty calibration at different stages in the fusion pipeline are presented and compared on the ScanNet dataset. We show that the most widely used Bayesian fusion strategy is among the worst calibrated, and propose a learned pipeline that combines fusion and calibration, GLFS, which achieves simultaneously higher accuracy and 3D map calibration while retaining real-time capability and adding only 525 learned parameters to the pipeline. We further illustrate the importance of map calibration on a downstream task by showing that incorporating proper semantic fusion to an indoor object search agent improves its success rates.**

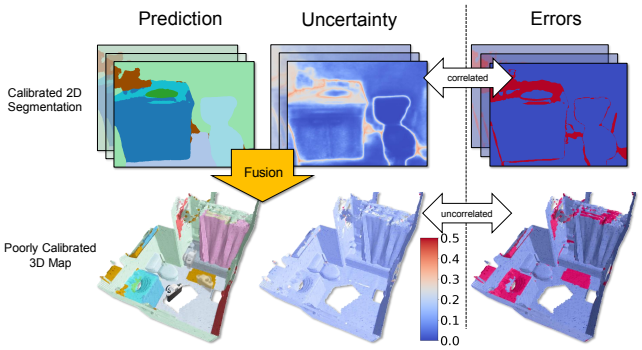


Fig. 1: The problem of fusion overconfidence: standard Recursive Bayesian semantic fusion [30] produces overconfident 3D semantic maps with low uncertainty (bottom) in error regions even when the 2D semantic segmentation images are well-calibrated (with uncertainty correlating to errors, see top). [Best viewed in color.]

I. INTRODUCTION

Semantic segmentation models output a distribution over a set of classes for each element in the scene, and the probability of the most likely class is called the model's **confidence**. The field of **confidence calibration** measures and aims to reduce the disparity between a model's confidence and its actual measured performance and has been widely studied in 2D vision tasks [1, 15, 27, 31, 60]. Calibrated uncertainty estimates can be crucial to decision making in many fields, such as autonomous driving [23, 42] and medical imaging [31, 33, 47], as it allows for more cautious action under uncertain conditions, improving safety and reliability. Semantic 3D mapping is an active research topic that seeks methods to build 3D maps in real-time while also leveraging image segmentation models from computer vision to label objects in the scene [30], and has numerous applications in robotics. However, the problem of confidence calibration in semantic 3D maps has yet to be addressed.

In this paper, we articulate a general **overconfidence problem** in semantic fusion calibration in which the semantic estimates in the 3D map develop significant overconfidence (Fig. 1). We highlight the unintuitive result that map overconfidence is not mitigated by the use of well-calibrated image segmentation, and is instead a result of the fusion strategy. Possible causes of this phenomenon include independence assumptions used in the fusion strategy, sensitivity to outliers, and viewpoint distribution biases. Moreover, we show the most widely used fusion strategy, Recursive Bayesian Update (RBU) [30], to be especially susceptible to overconfidence.

We identify several approaches to mitigate these problems and show that adopting alternative fusion strategies, like Naïve Averaging [29], tend to improve calibration. We introduce a method for calibrating the image segmentation model via Temperature / Vector Scaling [15] to directly optimize the calibration of final fused maps. Moreover, we present Generalized Learned Fusion Strategy (GLFS), an end-to-end trainable framework that learns logit Vector Scaling, sample weighing, and fusion method in a unified fashion, while remaining real-time capable. Experiments on the ScanNet dataset show that these strategies improve map calibration without degrading accuracy. Finally, we show the utility of better calibrated maps in a robotics application. In an object-goal-navigation task [6], better calibrated maps lead to performance improvements as the agent can better handle overconfident segmentation model outputs through leveraging diverse object viewpoints for reliable detection. Software for all methods and experiments is publicly available at *https://github.com/joaomcm/Semantic-3D-Mapping-Uncertainty-Calibration*.

II. RELATED WORK

Semantic mapping aims to create a labelled 3D map of the environment from a sequence of RGB-D images or depth readings. Individual frames are given to semantic segmentation models to produce pixel labels that are fused into estimates for each map element. Methods vary in their choice of map representation, such as voxels [2, 21, 29, 44, 46], surfels [30, 48], points [19, 51], Neural Radiance Fields [26] or sparse Gaussian Processes [61], and some

J. M. C. Marques was supported by NIFA Award #2021-67021-34418.
[1]: Department of Computer Science at the University of Illinois at Urbana-Champaign, Urbana, Illinois, USA.
(jmc12,azhai2,shenlong,kkhauser)@illinois.edu

methods identify objects as sub-elements of the map [14, 48, 51]. The overconfidence problem generally exists across map representations. Methods also vary in the choice of SLAM back-end, semantic segmentation model, and fusion strategy. Existing fusion strategies include Recursive Bayesian Update (RBU) [2, 5, 21, 30, 32, 35, 44, 46], histogram/weighed average-based strategies [6, 14, 29, 49, 52] and learning-based aggregation [13, 26, 53, 61]. Although most past work focuses on reconstruction accuracy or efficiency, overconfidence has also been noted [29, 32]. It has been addressed via assorted methods such as using averaging fusion [29] and sample weighing via epistemic uncertainty from Bayesian neural networks [32], but a systematic study of semantic map calibration, as we do here, has not yet been done.

Calibration is an important topic in deep learning due to observed model overconfidence using standard training methods [15, 25, 55]. Model calibration is most often assessed using reliability diagrams [37], which relates a model's confidence to the likelihood that it is correct. The most widely used metric is Expected Calibration Error (ECE) [15, 34], and a variety of post-hoc calibration methods have been used to recalibrate models including Temperature and Vector Scaling [15] and Platt scaling [43], with many more extensions aimed at better calibrating multi-class predictors [16, 24] or Detectors [27]. The literature has also identified many issues with ECE, in particular for multi-class classification [16, 39], like class frequency dependence, sensitivity to bin number, and non-differentiability [4]. Calibration has not yet been applied to semantic fusion because the "model" in this case is the output of a large set of image segmentation model outputs fused into a segmentation of a 3D map element (voxel). This paper proposes to measure and optimize the calibration of an entire semantic fusion pipeline, which requires simultaneous optimization across all views and voxels.

Maps with calibrated uncertainty are helpful in several applications. Past work has used semantic uncertainty as an objective for next-best-view problems for improved semantic reconstruction [12] or as information gain objectives for active map exploration [2]. We apply this to object goal navigation (ObjectNav), where an agent, placed in an unknown environment, navigates to a specific object (e.g., "chair") using posed RGB-D observations [3]. We show that improved calibration in semantic fusion increases the success rates of a recent ObjectNav agent [58] while using the same base segmentation model for semantics.

III. BACKGROUND

We begin with a brief summary of semantic 3D mapping, focusing on voxel representations [8, 10, 38]. Each entity in the voxel grid encodes geometry [38] (such as occupancy or signed distance field), appearance (e.g., colors), and semantic or instance labels [17, 54] (in the form of a categorical probability vector). Given a series of sensor observations, the **fusion** procedure aggregates the information from different viewpoints over time into each voxel.

The mapping algorithm (fusion) takes as input a series of RGB-D images and poses $\mathbf{I}^t = \{(\mathbf{C}^t, \mathbf{D}^t, \mathbf{T}^t)\}_{t=1,\dots,T}$, where $\mathbf{C}^t \in \mathbb{R}^{H\times W\times 3}$ is the RGB color, $\mathbf{D}^t \in \mathbb{R}^{H\times W}$ is the depth and $\mathbf{T}^t \in SE(3)$ is the pose generated by SLAM. The output of the metric reconstruction is a voxel grid $\mathbf{V} = \{\mathbf{v}_i = (\delta_i, w_i, \boldsymbol{\gamma}_i)\}_{i=1,\dots N}$ that encodes the geometry of 3D space. Each voxel $\mathbf{v}_i$ represents geometry attributes at a 3D location and contains three values: a signed distance to the nearest surface δ_i, a weight w_i used for averaging, and, optionally, a color tuple γ_i.

At every time t, we perform metric reconstruction following the procedure described in Nießner *et al.* [38] and obtain the reconstructed surface mesh using the marching cubes algorithm [28].

A. Semantic Fusion

Given the geometric representation, **semantic fusion** additionally assigns a semantic label l_i to each voxel $\mathbf{v}_i$. To achieve this, semantic information is extracted from 2D image observations across viewpoints and integrated over visible voxels. A class probability vector $\mathbf{s}_i \in \mathbb{R}^K$ is associated with the voxel $\mathbf{v}_i$ where the k-th element represents the probability that the voxel belongs to the semantic category k: $s_{ik} = P(l_i = k|\mathbf{I}_{1,\dots,T})$, ensuring $\sum_k s_{ik} = 1$. Each incoming image $\mathbf{I}^t$ is segmented to generate a semantic map $\mathbf{S}^t \in \mathbb{R}^{W\times H\times K}$. Then, for each voxel $\mathbf{v}_i$ visible on frame t, we project it into image coordinates and query its semantic class vector s^t_{ik} (accounting for visibility). This value is designated as the likelihood probability: $P(l_i = k|\mathbf{I}^t) = \mathbf{s}^t_i$. We wish to model the posterior of the voxel label given these observations: $P(l_i = k|\mathbf{I}_{1,\dots,T})$.

McCormac *et al.* [30] proposed the **Recursive Bayesian Update** (RBU) formula, which assumes independence between observations and approximates the posterior through Bayes's Rule as:

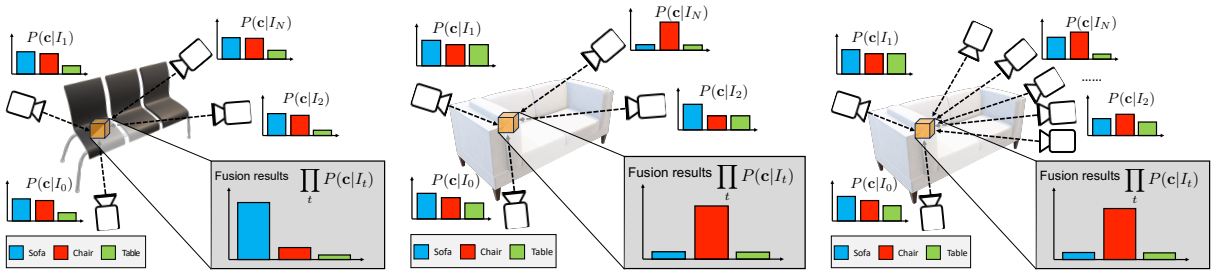
$$\mathbf{s}_i = P(l_i|\mathbf{I}_{1,\dots,T}) \propto P(l_i) \prod_{t\in F_i} \mathbf{s}^t_i, \tag{1}$$

$P(l_i)$ representing a prior distribution and F_i the set of frames in which $\mathbf{v}_i$ is observed. Multiplication is performed element-wise. Assuming a uniform prior, $\mathbf{s}_i = \frac{1}{Z}\prod_t \mathbf{s}^t_i$, where Z normalizes $\mathbf{s}_i$ into a valid probability vector. This approximation can be recursively estimated via $\mathbf{s}_i \leftarrow \frac{1}{Z}\mathbf{s}_i\mathbf{s}^t_i$. In practice, this update is done in log space after a Laplace smoothing of $\mathbf{s}^t_i$ for numerical stability. Although other fusion methods exist (Sec. V-A), RBU remains the most widely used thanks to its simplicity and probabilistic interpretation.

B. Measuring Map Calibration

We now discuss and present a few ways to measure 3D mapping miscalibration. Let the ground truth label of a voxel $\mathbf{v}_i$ be l^*_i, and its estimated posterior of belonging to class k, $\mathbf{s}_{ik}$. Let π_i denote the top-1 label of voxel $\mathbf{v}_i$, $\pi_i = \arg\max_k(\mathbf{s}_{ik})$, and h_i denote its predicted confidence, $h_i = \max_k(\mathbf{s}_{ik})$. Further, let there be N total labelled voxels.

The most widely used metric is Expected Calibration Error (ECE) [15, 34]. Partition the interval [0,1] in O uniformly-spaced bins B_b and match each sample to a bin by its $h(\mathbf{v}_i)$ value. Define bin *confidence* as the mean



(a) Aggregation overconfidence (b) Sensitivity to outliers (c) Viewpoint sampling biases

Fig. 2: Illustrating potential causes of fusion overconfidence. [Best viewed in color.]

value of the model's confidence for samples within the bin $\texttt{Conf}(B_b) = \frac{\sum_{h_i \in B_b} h_i}{|B_b|}$ and define the mean bin accuracy $\texttt{Acc}(B_b) = \frac{\sum_{h_i \in B_b} \mathbb{1}[\pi_i = l_i^*]}{|B_b|}$. The ECE is then the weighted difference between bin accuracy and confidence, $\texttt{ECE} = \sum_{b=1}^{O} \frac{|B_b|}{N} |\Delta(B_b)|$ where $\Delta(\cdot) = \texttt{Acc}(\cdot) - \texttt{Conf}(\cdot)$.

ECE is ill-suited for evaluating multi-class calibration since it is skewed by performance on common classes [16, 39], and in 3D mapping, classes may have orders of magnitude more voxels than others. Many alternatives have been proposed to address these flaws [16, 24, 39]. Top-Label ECE (TL-ECE) [16] performs binning conditioned on the predicted classes. It associates a sample with bin B_{bk} if $h(\mathbf{v}_i) \in B_b$ and $\pi_i = k$, and calculates a bin-weighted error:

$$\texttt{TL-ECE} = \sum_{b=1}^{O} \sum_{k=1}^{K} \frac{|B_{bk}|}{N} |\Delta(B_{bk})| \tag{2}$$

However, this metric is still heavily influenced by class frequency, since it weighs contributions to the metric by $\frac{|B_{bk}|}{N}$. We provide in this paper an alternative measure of calibration, which is agnostic to class frequency, mean Expected Calibration Error (mECE). Similarly to Class-Conditional ECE [24], it is obtained by calculating a class-conditional ECE for every class, but unlike Kull *et al.* [24], that bin predictions based on $h(\mathbf{v}_i)$ and π_i, we bin predictions based on $h(\mathbf{v}_i)$ and the *ground truth class*, $l*$, and take the average value across classes. This avoids the calibration failure case of a predictor obtaining perfect calibration by always predicting the most common class with the confidence given by its frequency. Thus, binning B_{bk^*} is conditioned on $l^* = k$, and mECE is defined as:

$$\texttt{mECE} = \frac{1}{K} \sum_{k^*=1}^{K} \sum_{b=1}^{O} \frac{|B_{bk^*}|}{\sum_{b=1}^{O} |B_{bk^*}|} |\Delta(B_{bk^*})|, \tag{3}$$

retaining the notion of $\Delta(\cdot)$ from before.

We argue that this metric better captures the overall behavior of the 3D calibration model across all relevant classes, since it weights classes equally, i.e., miscalibration of a class' bins is scaled relative to their own class's frequency. The high-level difference between mECE and TL-ECE is analogous to the difference between mIoU and f-mIoU - i.e. mECE avoids drowning out minority classes.

IV. THE FUSION OVERCONFIDENCE PROBLEM

Accuracy and **calibration** are the two principal desiderata for semantic 3D mapping. If we cannot guarantee 100% accuracy, we at least expect our predictions to be well-calibrated. For instance, if we predict that a voxel is a sofa with a 60% probability, denoted as $s_{ik} = 0.60$, then this prediction should be correct about 60% of the time. If a model consistently gives confidence scores that exceed its actual performance, it is deemed **overconfident**. Such overconfidence is often observed in 2D image-based perception tasks, and various uncertainty calibration strategies have been employed to address it [16, 24, 27].

However, calibration in 3D semantic mapping remains unaddressed. As briefly noted by prior work [29], The Recursive Bayesian Update [30] in Eq. 1 tends to produce highly overconfident 3D semantic maps. We further illustrate this phenomenon in Fig. 1. Note that despite the fact that the 2D semantic cues exhibit only mild miscalibration, the resulting fused map using RBU displays severe overconfidence, or worse disparity between accuracy and confidence.

Fusion overconfidence can have a few different causes, some of which are illustrated in Figure 2:

Uncalibrated 2D Segmentation: Many image-based semantic segmentation models are poorly calibrated [27]. In particular, models trained with cross-entropy loss often demonstrate overconfident behavior [15]. Integrating overconfident estimates with RBU will result in an overconfident 3D prediction due to its sensitivity to outliers.

Incorrect Independence Assumption: RBU assumes that observation likelihoods are independent, overlooking dependencies across views and 2D model biases. These biases exist in 2D deep models due to a variety of reasons, such as dataset collection biases [27], leading to a "double-counting" effect (Figure 2a). For instance, if a calibrated model consistently predicted $\mathbf{s} = (0.49, 0.51)$ across 50 neighboring viewpoints, RBU would predict class 2 with 88% confidence, even though each observation contributes little evidence overall.

Sensitivity to Outliers: As pointed by Morilla-Cabello *et al.* [32], a single overconfident input can drastically change the RBU posterior, as in Figure 2b. Laplace Smoothing of semantic likelihoods mitigates but does not fully solve this.

Viewpoint Coverage Dependence - Most data collection in the real world cannot guarantee a well-covered, uniform sampling of viewpoints. This inevitably results in a lack of dense coverage, e.g., one might only see the sofa from a certain angle. If deep models are sensitive to viewpoints [27], as in Figure 2c, this can lead to confident errors.

Non-Uniform Viewpoint Density - Similarly, if the camera trajectory lingers longer in one region than another, as shown

TABLE I: **Quantitative Analysis of Logit Scaling**: 2D calibration is insufficient to produce well-calibrated 3D semantic maps.

Model	Calibration	Pixel mECE ↓	Voxel mECE↓
Segformer	None	0.124	0.451
	2D Temperature	**0.114**	0.450
	3D Temperature	0.546	**0.176**
ESANet	None	0.187	0.292
	2D Temperature	**0.159**	0.294
	3D Temperature	0.656	**0.177**

in Figure 2c, positional biases in the model can steer it towards overconfidence, despite good pose diversity.

Metric Mapping Errors - Errors in geometric map reconstruction, whether from depth sensing errors or uncertainties in metric reconstruction, can result in spurious 2D-3D associations, leading to overconfident or inconsitent predictions.

Notably, 2D overconfidence is only a small contributor to fusion overconfidence. We calibrate two semantic segmentation models: Segformer [54] and ESANET [50] with Temperature Scaling on 21 scenes from ScanNet [9]. We then reconstruct another set of 100 scenes using RBU fusion for both uncalibrated and calibrated models. Table I shows the mECE of the RGBD image pixels and the mECE of the voxels of the fused map. Traditional 2D calibration improves **pixel** mECE but **voxel** mECE is barely changed. In contrast, we will introduce a 3D temperature scaling method for calibrating voxel mECE directly, leading to maps that are much better calibrated. Surprisingly, segmentation models that are calibrated to yield better calibrated maps can be quite uncalibrated at the pixel level!

V. METHODS

We identify and discuss three classes of approaches to address fusion overconfidence in the context of real-time semantic mapping: **alternative fusion strategies**, **down-weighting or filtering samples** and **calibrating the semantic segmentation model.** Finally, we also present a generalized method that simultaneously learns the fusion, weighting, and image segmentation calibration to optimize 3D calibration metrics without harming model accuracy.

A. Other Fusion Strategies

Alternative fusion strategies have been proposed to address some of the limitations of RBU. The **Histogram** fusion approach considers each image label as a "semantic vote" in favor of a class. The element probability is then given by:

$$\mathbf{s}_i \propto \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}[k = \arg\max s_i^t] \tag{4}$$

(For brevity, treat the voxel as being visible at all $t = 1, \ldots, T$). This strategy is widely used in ObjectNav [7, 45], panoptic segmentation [14] and outdoor reconstruction [52].

The **Naïve Averaging** idea treats the image segmentation likelihoods as fractional votes for each class [29, 49, 52]:

$$\mathbf{s}_i \propto \frac{1}{T}\sum_{t=1}^{T} \mathbf{s}_i^t \tag{5}$$

Both histogram voting and averaging voting integrate the semantic logits additively, which compromises their probabilistic interpretation. However, they are less sensitive to overconfident prediction outliers.

A final alternative we consider is the **Geometric Mean** of likelihoods, as a logical extension of Naïve Averaging

$$\mathbf{s}_i \propto \left(\prod_{t=1}^{T} \mathbf{s}_i^t\right)^{\frac{1}{T}}, \tag{6}$$

which is also performed in log space after Laplace smoothing of $\mathbf{s}_i^t$ for numeric stability. The geometric mean has the advantage of being less sensitive to outliers and is less prone to "double-counting" compared to RBU.

Other alternatives exist, like max-fusion [59] and deep learned fusion [53, 57], but are likely to be more susceptible to outliers [59] or not real-time capable [53, 57].

B. Sample Weighting

Since semantic segmentation models and depth sensors can exhibit distance and pose related biases [27], one might not wish to give all observations equal weights in fusion [8]. As such, each of the integration strategies above can be modified to use weights to capture some degree of confidence in each pixel's contribution to the 3D element *in dimensions orthogonal to the image segmentation confidence*. For example, Curless and Levoy [8] propose to weigh samples based on the estimated normal and distance to the camera, Küppers *et al.* [27] proposes image-position-based calibration of uncertainties, and Qiu *et al.* [44] propose a quadratic sample weight based on voxel distance to camera to reflect a "segmentation distance sweet-spot". Recent work proposes to weigh samples based on estimated epistemic uncertainty using Monte-Carlo dropout on segmentation networks [32], which is ill-suited for real-time applications.

C. Calibrating Image Segmentation via Logit Scaling

The final category of approaches we consider is to calibrate the confidence in image segmentation likelihoods $\mathbf{s}_i^t$. In this paper, we consider two baseline calibration methods, Temperature and Vector Scaling, known to work well in 2D semantic tasks [15]. Let the image semantic likelihoods be $\mathbf{S}^t = \sigma_{SM}(\lambda^t)$, where λ^t are the segmentation model logits and σ_{SM} is the softmax function. Define a temperature parameter τ and a calibrated likelihood as $\mathbf{S}^t(\tau) = \sigma_{SM}\left(\frac{\lambda^t}{\tau}\right)$. Given a calibration metric Ω such as mECE, 2D temperature scaling minimizes calibration error of all images over τ:

$$\tau = \arg\min_{\tau} \Omega\left(\{\mathbf{S}_{u,v}^t(\tau) \quad \forall u, v, t\}\right). \tag{7}$$

Vector scaling uses a per-class scaling vector $\tau \in \mathbb{R}^K$ and scales the logits classwise as $\mathbf{S}_{uv}^t(\tau) = \sigma_{SM}\left(\left\{\frac{\lambda_{u,v,k}^t}{\tau_k} \quad \forall k = 1, \ldots, K\right\}\right)$.

This paper introduces a new concept of *3D Temperature / Vector scaling*. The calibration metric is no longer assessed over images but rather over the fused voxels of the map. For each voxel $\mathbf{v}_i$, its estimated posterior is a function of τ, $\mathbf{s}_i(\tau) = \texttt{Fusion}(\mathbf{s}_i^1(\tau), \ldots, \mathbf{s}_i^T(\tau))$, where $\texttt{Fusion}(\ldots)$ is any of the described methods in Section V-A. This can then similarly be optimized in vector and temperature variants:

$$\tau = \arg\min_{\tau} \Omega(\{\mathbf{s}_i(\tau) \quad \forall \mathbf{v}_i \in \mathbf{V}\}) \tag{8}$$

Logit scaling is usually performed with multiple image sequences, seeking to minimize the mean calibration error.

D. Generalized Learned Fusion Strategy (GLFS)

Finding the optimal mix of fusion, sampling, and logit scaling strategies is cumbersome. We address this challenge by introducing a generalized learned fusion strategy, GLFS. The key idea is to use differentiable gating variables to switch between various integration strategies and to make both the temperature and sample weighting learnable as in Eq. 9

$$\mathbf{s}_i \propto G\ e^{\left(\Sigma_t \alpha \left(w_i^t ln(\mathbf{s}_i^t(\tau))\right)\right)} + (1-G)\Sigma_t \alpha \left(w_i^t \mathbf{s}_i^t(\tau)\right) \quad (9)$$

where $\alpha = \left(\frac{(1-\epsilon)}{\Sigma_t w_i^t} + \epsilon\right)$ is a gating variable that interpolates between RBU and Geometric Mean, while G is a gating variable that balances between the arithmetic mean and the geometric mean for integration. w_i^t weighs the importance of each pixel contributing to the final voxel semantics and τ represents the logit scaling vector.

This unified procedure generalizes to all the listed fusion strategies and incorporates logit adjustments via vector scaling and pose-dependence adjustments through learned sample weighting. When $\tau = 1$ and $w_i^t = 1\ \forall i, t$, it reduces to RBU, Geometric Mean, and Naïve Averaging when $(G, \epsilon) = (1, 1)$, $(1, 0)$, $(0, 0)$, respectively. Similarly, it reduces to histogram fusion when $|\tau| \to 0$ and $(G, \epsilon) = (0, 0)$. Thus, this model is capable of capturing the behavior of all fusion strategies we have detailed.

To account for pose biases in segmentation models, w_i^t comes from a learned look-up-table, $\boldsymbol{M}$, which takes as entries π_i^t (i.e, the post-scaling predicted class of the pixel associated with this voxel at this time), d_i^t, the distance from this voxel to the camera, and α_i^t, the incidence angle between the camera ray and the current estimate of the surface normal at that voxel, hopefully capturing the effects of the weighing heuristics through learning. We define $\theta = (\tau, G, \epsilon, \mathbf{M})$, and the calibration error is defined as $\texttt{mECE}(\{\mathbf{s}_i(\theta)\,\forall i\})$.

However, mECE is not directly end-to-end trainable because it is not differentiable, so we use a differentiable analogue, DECE [4], for which we can also define its mECE equivalent, which we call mDECE. Our overall loss function balances calibration with accuracy:

$$\mathcal{L} = \eta\ \texttt{mDECE}(\mathbf{s}_i(\theta), l_i^*) + \texttt{NLL}(\mathbf{s}_i(\theta), li^*) \quad (10)$$

where η scales the calibration term relative to the accuracy term (Negative Log Likelihood). This loss simultaneously promotes uncertainty calibration while minimizing the compromise in prediction accuracy. The calibration model parameters are then learned through backpropagation with PyTorch [41], leveraging voxel caching for efficiency.

VI. EXPERIMENTS

A. Segmentation Models and Datasets

To evaluate the effect of different strategies on 3D fusion calibration, we perform reconstruction experiments on the ScanNet Dataset [9] using two different semantic segmentation models: Segformer [54] and ESANet [50]. These are representative segmentation networks for RGB and RGB-D, respectively. Moreover, they utilize two differing backbone architectures: Vision Transformers [11] and Convolutional Neural Networks, [18]. to help us evaluate the generality of the various calibration strategies. Segformer was pretrained on the ADE-20K dataset using the b4 backbone [20] and ESANet was pretrained on the NYU_v2 dataset [36]. Both models were then finetuned on ScanNet's 20 classes on a subset of every 30-th frame of the train split recordings of the ScanNet dataset, with their classification head trained from scratch. We implement mapping strategies using Open3D's real-time GPU-based dense SLAM implementation [10].

We divided the scans from the ScanNet V2 validation set into calibration training (79 scans), calibration validation (21 scans), and calibration testing (100 scans). Since there are multiple scans for the same scene, we make sure there is no scene overlap between testing and validation/training sets.

B. Mapping Calibration Experiments

We compare the calibration and accuracy of the fusion strategies and 3D calibration methods described in Sec. III.

Bayesian optimization [40] is used for temperature and vector scaling, using the mECE metric and Upper Confidence Bound (UCB) acquisition function with $\beta = 1$, the Mattern 2.5 kernel and 1000 maximum samples. The models were initialized with a sweep of 50 temperatures in log space between the maximum and minimum temperatures for temperature calibration (0.01, 200). Then, we narrowed the search for the vector scaling parameters to be closer to the optimal temperature scaling parameter, allowing temperatures to be within a 50% difference of each optimized value. A similar log-scale diagonal sweep is performed as initialization for the vector parameters - as well as 30 random samples.

Results are reported on the calibration test split using ground-truth poses. Figure 3 plots calibration (mECE, TL-ECE, and Brier Scores) vs accuracy (mIoU). We observe a clear tradeoff between accuracy and calibration. Temperature scaling tends to somewhat improve calibration with minimal degradation in accuracy, but vector scaling is often highly detrimental to accuracy for sometimes modest calibration improvements. Averaging-based fusion is generally a strong performer. Our learned model typically exceeds other methods in either accuracy or calibration, improving calibration without degrading accuracy, as intended. We also found that the ESANet model proved harder to calibrate. GLFS parameter optimization was achieved in under 30 hours on an NVidia RTX A40 needing less than 15 GB of VRAM and GLFS reconstruction was at most 15% slower than baseline reconstruction algorithms due to normal estimation overhead.

Qualitative results comparing RBU to our learned method are shown in Fig. 4. The error-uncertainty disparity for RBU is high, demonstrating severe overconfidence. In contrast, GLFS is less confident in its mistakes, improving mECE.

C. Object Goal Navigation Experiments

We integrate our semantic fusion pipelines with a recent modular Object Goal Navigation agent, PEANUT [58]. We use our live metric-semantic reconstruction pipeline to generate a 3D semantic point cloud with different fusion strategies, which is then projected into the 2D plane and overlaid over PEANUT's obstacle and exploration map. During the projection, points get binned based on their ground coordinates

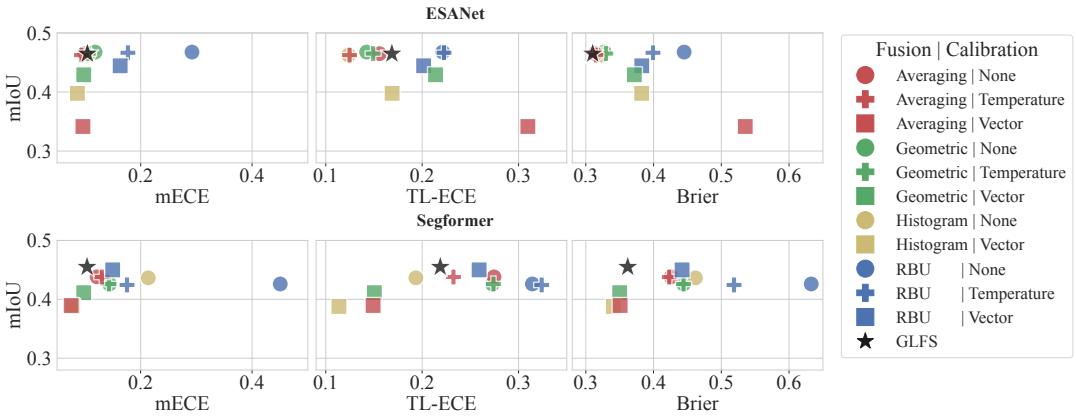


Fig. 3: Map calibration vs. accuracy for combinations of fusion and calibration strategies on 100 ScanNet scenes. Higher is better for mIOU; lower is better for mECE, TL-ECE, and Brier - i.e. better models are always closer to the upper left corner. [Best viewed in color.]

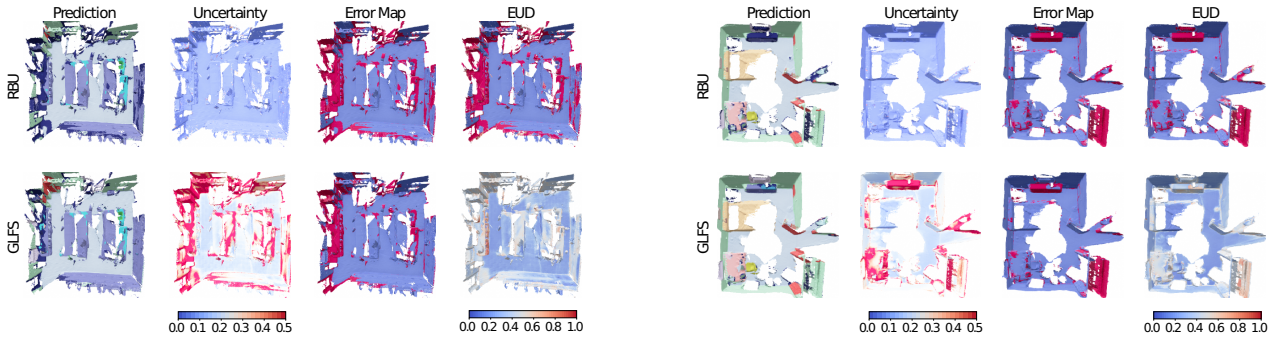


Fig. 4: Qualitative comparison of RBU (top) and our proposed GLFS method (bottom) on two ScanNet scenes reconstructed using the Segformer model. Although similarly accurate, GLFS provides significantly improved calibration. EUD = Error-Uncertainty Disparity [Best viewed in color.]

and on π_i for each point. Then, the average of the point confidence predictions within that (x, y, l) bin represents the confidence of the class l being in (x, y) on the 2D map. This semantically fused map is fed to PEANUT's goal prediction network **without finetuning it** to the new distribution of semantic maps due to time constraints. Besides the semantic fusion, the only major difference between our version and PEANUT is that we perform detection by thresholding the fused 2D semantic map, i.e., on the projected $\mathbf{s}_i$, while PEANUT's detection of an object is performed by thresholding an image-level confidence through a Mask-RCNN model [17]. Finally, we apply connected component analysis to the goal map, retaining only the largest goal connected component to account for voxel misprojections. We finetune a Segformer [54] model on a set of training images from random exploration on HM3D v0.1 [56] as our agent's semantic segmentation. We evaluate the performance of the agent under 3 different conditions: On vanilla PEANUT, by thresholding Segformer's semantic mask in image space, and on our semantically fused implementation with both RBU (Fusion-RBU) and Naïve Averaging (Fusion-NA) fusion. For each mixed model, we tune PEANUT's hyperparameters, like collision radius and semantic detection threshold, on the first 500 validation episodes of the 2022 Habitat ObjectNav challenge, and evaluate it on episodes 500-999 of the same challenge. ObjectNav metrics [3] of these agents and a ground truth segmentation ablation, presented in Table II, show that the empirically better calibrated 3D Naïve Averaging fusion can result in higher agent success rates when compared to Chaplot *et al.* [7]'s Neural Deterministic Fusion (NDF) or the overconfident RBU, at the expense of requiring more evidence to detect target objects, resulting in slightly worse SPL.

TABLE II: Object goal navigation performance on HM3D-val for various variations of PEANUT and (standard error of the mean). The calibration strategies studied in this paper generally improve performance.

Map Type	Segmentation	SPL ↑	Success Rate ↑	SSPL ↑	DTG ↓
NDF [6]	**GT**	0.384 (0.0132)	0.724 (0.0200)	0.401 (0.0123)	2.876 (0.2534)
NDF [6]	Segformer	0.295 (0.0130)	0.596 (0.0220)	0.328 (0.0121)	3.764 (0.2639)
Fusion-NA	Segformer	0.312 (0.0132)	**0.604** (0.0219)	0.347 (0.0120)	3.576 (0.2589)
Fusion-RBU	Segformer	**0.314** (0.0134)	0.596 (0.0220)	**0.351** (0.0123)	**3.549** (0.2595)

VII. CONCLUSIONS AND FUTURE WORK

We introduce the problem of semantic calibration in real-time capable metric-semantic mapping pipelines and show that better calibration can be achieved through an end-to-end unified approach to semantic fusion, GLFS, without sacrificing segmentation performance. We also show that better semantic fusion improves the performance of modular ObjectNav agents when using the same semantic segmentation model by providing robustness to outlying predictions.

Three avenues are considered for future work: First, we want to quantify the effects of other calibration strategies like position-dependent calibration [27] and Dirichlet calibration [24] on 3D semantic map calibration, and novel 3D specific calibration procedures. Second, we wish to investigate how ObjectNav agents can better leverage 3D semantic uncertainty to guide their exploration [2] and how that would affect their performance. Finally, we would like to study the quantification and improvement of uncertainty in open-set metric-semantic maps [12, 22].

REFERENCES

[1] I. Abaspur Kazerouni, L. Fitzgerald, G. Dooly, and D. Toal, "A survey of state-of-the-art on visual SLAM," *Expert Systems with Applications*, vol. 205, p. 117 734, 2022.

[2] A. Asgharivaskasi and N. Atanasov, *Active Bayesian Multi-class Mapping from Range and Semantic Segmentation Observation*, 2021.

[3] D. Batra, A. Gokaslan, A. Kembhavi, O. Maksymets, R. Mottaghi, M. Savva, A. Toshev, and E. Wijmans, *ObjectNav Revisited: On Evaluation of Embodied Agents Navigating to Objects*, 2020.

[4] O. Bohdal, Y. Yang, and T. Hospedales, "Meta-Calibration: Learning of Model Calibration Using Differentiable Expected Calibration Error," Jun. 2021.

[5] S. Bultmann, J. Quenzel, and S. Behnke, "Real-time multi-modal semantic fusion on unmanned aerial vehicles with label propagation for cross-domain adaptation," *Robotics and Autonomous Systems*, vol. 159, p. 104 286, 2023.

[6] D. S. Chaplot, D. Gandhi, S. Gupta, A. Gupta, and R. Salakhutdinov, "Learning to explore using active neural slam," *arXiv preprint arXiv:2004.05155*, 2020.

[7] D. S. Chaplot, D. P. Gandhi, A. Gupta, and R. R. Salakhutdinov, "Object goal navigation using goal-oriented semantic exploration," in *NeurIPS*, 2020.

[8] B. Curless and M. Levoy, "A volumetric method for building complex models from range images," in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, 1996, pp. 303–312.

[9] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Niessner, "ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes," in *CVPR*, Jul. 2017.

[10] W. Dong, Y. Lao, M. Kaess, and V. Koltun, "ASH: A Modern Framework for Parallel Spatial Hashing in 3D Perception," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–18, 2022.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *ICLR*, 2021.

[12] F. Engelmann, F. Manhardt, M. Niemeyer, K. Tateno, M. Pollefeys, and F. Tombari, "Open-Set 3D Scene Segmentation with Rendered Novel Views," *arXiv*, 2023.

[13] X. Fu, S. Zhang, T. Chen, Y. Lu, L. Zhu, X. Zhou, A. Geiger, and Y. Liao, "Panoptic NeRF: 3D-to-2D Label Transfer for Panoptic Urban Scene Segmentation," Mar. 2022.

[14] M. Grinvald, F. Furrer, T. Novkovic, J. J. Chung, C. Cadena, R. Siegwart, and J. Nieto, "Volumetric Instance-Aware Semantic Mapping and 3D Object Discovery," *IEEE RA-L*, vol. 4, no. 3, pp. 3037–3044, 2019.

[15] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in *ICML*, vol. 70, 2017, pp. 1321–1330.

[16] C. Gupta and A. Ramdas, "Top-label calibration and multiclass-to-binary reductions," in *ICLR*, 2022.

[17] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *ICCV*, Oct. 2017.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[19] A. Hermans, G. Floros, and B. Leibe, "Dense 3D semantic mapping of indoor scenes from RGB-D images," in *ICRA*, 2014, pp. 2631–2638.

[20] *https://huggingface.co/nvidia/segformer-b4-finetuned-ade-512-512*.

[21] N. Hughes, Y. Chang, and L. Carlone, "Hydra: A Real-time Spatial Perception Engine for 3D Scene Graph Construction and Optimization," *arXiv*, Jan. 2022.

[22] K. M. Jatavallabhula, A. Kuwajerwala, Q. Gu, M. Omama, T. Chen, S. Li, G. Iyer, S. Saryazdi, N. Keetha, A. Tewari, J. B. Tenenbaum, C. M. de Melo, M. Krishna, L. Paull, F. Shkurti, and A. Torralba, *ConceptFusion: Open-set Multimodal 3D Mapping*, 2023.

[23] A. Kendall and Y. Gal, "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?" In *NeurIPS*, vol. 30, 2017.

[24] M. Kull, M. Perello Nieto, M. Kängsepp, T. Silva Filho, H. Song, and P. Flach, "Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with Dirichlet calibration," in *NeurIPS*, vol. 32, 2019.

[25] A. Kumar, P. S. Liang, and T. Ma, "Verified Uncertainty Calibration," in *NeurIPS*, vol. 32, 2019.

[26] A. Kundu, K. Genova, X. Yin, A. Fathi, C. Pantofaru, L. J. Guibas, A. Tagliasacchi, F. Dellaert, and T. Funkhouser, "Panoptic Neural Fields: A Semantic Object-Aware Neural Scene Representation," in *CVPR*, Jun. 2022, pp. 12 871–12 881.

[27] F. Küppers, A. Haselhoff, J. Kronenberger, and J. Schneider, "Confidence calibration for object detection and segmentation," in *Deep Neural Networks and Data for Automated Driving: Robustness, Uncertainty Quantification, and Insights Towards Safety*, Springer International Publishing Cham, 2022, pp. 225–250.

[28] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3D surface construction algorithm," *ACM SIGGRAPH Computer Graphics*, vol. 21, no. 4, pp. 163–169, 1987.

[29] J. Mccormac, R. Clark, M. Bloesch, A. Davison, and S. Leutenegger, "Fusion++: Volumetric Object-Level SLAM," in *International Conference on 3D Vision (3DV)*, 2018, pp. 32–41.

[30] J. McCormac, A. Handa, A. Davison, and S. Leutenegger, "SemanticFusion: Dense 3D semantic mapping with convolutional neural networks," in *ICRA*, 2017, pp. 4628–4635.

[31] A. Mehrtash, W. M. Wells, C. M. Tempany, P. Abolmaesumi, and T. Kapur, "Confidence Calibration and Predictive Uncertainty Estimation for Deep Medical Image Segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 12, pp. 3868–3878, 2020.

[32] D. Morilla-Cabello, L. Mur-Labadia, R. Martinez-Cantin, and E. Montijano, "Robust Fusion for Bayesian Semantic Mapping," *arXiv*, Mar. 2023.

[33] B. Murugesan, B. Liu, A. Galdran, I. B. Ayed, and J. Dolz, "Calibrating segmentation networks with margin-based label smoothing," *Medical Image Analysis*, vol. 87, p. 102 826, 2023.

[34] M. P. Naeini, G. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using bayesian binning," in *AAAI*, 2015.

[35] G. Narita, T. Seno, T. Ishikawa, and Y. Kaji, "PanopticFusion: Online Volumetric Semantic Mapping at the Level of Stuff and Things," in *IROS*, 2019, pp. 4205–4212.

[36] P. K. Nathan Silberman Derek Hoiem and R. Fergus, "Indoor Segmentation and Support Inference from RGBD Images," in *ECCV*, 2012.

[37] A. Niculescu-Mizil and R. Caruana, "Predicting Good Probabilities with Supervised Learning," in *ICML*, 2005, 625–632.

[38] M. Nießner, M. Zollhöfer, S. Izadi, and M. Stamminger, "Real-time 3D reconstruction at scale using voxel hashing," *ACM ToG*, vol. 32, no. 6, pp. 1–11, 2013.

[39] J. Nixon, M. Dusenberry, G. Jerfel, T. Nguyen, J. Liu, L. Zhang, and D. Tran, "Measuring Calibration in Deep Learning," *Arxiv*, Apr. 2019.

[40] F. Nogueira, {*Bayesian Optimization*}*: Open source constrained global optimization tool for* {*Python*}.

[41] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems 32*, H Wallach, H Larochelle, A Beygelzimer, F d\textquotesingle Alché-Buc, E Fox, and R Garnett, Eds., Curran Associates, Inc., 2019, pp. 8024–8035.

[42] L. Peng, H. Wang, and J. Li, "Uncertainty Evaluation of Object Detection Algorithms for Autonomous Vehicles," *Automotive Innovation*, vol. 4, no. 3, pp. 241–252, 2021.

[43] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[44] R.-Z. Qiu, Y. Sun, J. M. Correia Marques, and K. Hauser, "Real-time Semantic 3D Reconstruction for High- Touch Surface Recognition for Robotic Disinfection," in *IROS*, 2022, pp. 9919–9925.

[45] S. K. Ramakrishnan, D. S. Chaplot, Z. Al-Halah, J. Malik, and K. Grauman, "Poni: Potential functions for objectgoal navigation with interaction-free learning," in *CVPR*, 2022.

[46] A. Rosinol, A. Violette, M. Abate, N. Hughes, Y. Chang, J. Shi, A. Gupta, and L. Carlone, "Kimera: From SLAM to spatial perception with 3D dynamic scene graphs," *The International Journal of Robotics Research*, vol. 40, no. 12-14, pp. 1510–1546, 2021.

[47] A.-J. Rousseau, T. Becker, J. Bertels, M. B. Blaschko, and D. Valkenborg, "Post Training Uncertainty Calibration Of Deep Networks For Medical Image Segmentation," in *International Symposium on Biomedical Imaging*, 2021, pp. 1052–1056.

[48] M. Runz, M. Buffier, and L. Agapito, "MaskFusion: Real-Time Recognition, Tracking and Reconstruction of Multiple Moving Objects," in *International Symposium on Mixed and Augmented Reality*, 2018, pp. 10–20.
[49] L. Schmid, J. Delmerico, J. L. Schönberger, J. Nieto, M. Pollefeys, R. Siegwart, and C. Cadena, "Panoptic Multi-TSDFs: a Flexible Representation for Online Multi-resolution Volumetric Mapping and Long-term Dynamic Scene Consistency," in *ICRA*, 2022, pp. 8018–8024.
[50] D. Seichter, M. Kohler, B. Lewandowski, T. Wengefeld, and H.-M. Gross, "Efficient RGB-D Semantic Segmentation for Indoor Scene Analysis," in *ICRA*, 2021, pp. 13 525–13 531.
[51] N. Sünderhauf, T. T. Pham, Y. Latif, M. Milford, and I. Reid, "Meaningful maps with object-oriented semantic mapping," in *IROS*, 2017, pp. 5079–5085.
[52] V. Vineet, O. Miksik, M. Lidegaard, M. Nießner, S. Golodetz, V. A. Prisacariu, O. Kähler, D. W. Murray, S. Izadi, P. Pérez, and P. H. S. Torr, "Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction," in *ICRA*, 2015, pp. 75–82.
[53] Y. Xiang and D. Fox, "DA-RNN: Semantic mapping with data associated recurrent neural networks," *arXiv preprint arXiv:1703.03098*, 2017.
[54] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, *SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers*, 2021.
[55] Z. Xiong, A. Eldesokey, J. Johnander, B. Wandt, and P.-E. Forssen, *Hinge-Wasserstein: Mitigating Overconfidence in Regression by Classification*, 2023.
[56] K. Yadav, R. Ramrakhya, S. K. Ramakrishnan, T. Gervet, J. Turner, A. Gokaslan, N. Maestre, A. X. Chang, D. Batra, M. Savva, A. W. Clegg, and D. S. Chaplot, *Habitat-Matterport 3D Semantics Dataset*, 2022.
[57] Z. Yang and C. Liu, "TUPPer-Map: Temporal and Unified Panoptic Perception for 3D Metric-Semantic Mapping," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 1094–1101.
[58] A. J. Zhai and S. Wang, "PEANUT: Predicting and navigating to unseen targets," in *ICCV*, 2023.
[59] J. Zhang, L. Dai, F. Meng, Q. Fan, X. Chen, K. Xu, and H. Wang, "3d-aware object goal navigation via simultaneous exploration and identification," in *CVPR*, 2023.
[60] J. Zhang, Y. Dai, M. Xiang, D.-P. Fan, P. Moghadam, M. He, C. Walder, K. Zhang, M. Harandi, and N. Barnes, "Dense Uncertainty Estimation," Oct. 2021.
[61] E. Zobeidi, A. Koppel, and N. Atanasov, "Dense Incremental Metric-Semantic Mapping via Sparse Gaussian Process Regression," in *IROS*, 2020, pp. 6180–6187.