

# EvidMTL: Evidential Multi-Task Learning for Uncertainty-Aware Semantic Surface Mapping from Monocular RGB Images

Rohit Menon

Nils Dengler

Sicong Pan

Gokul Krishna Chenchani

Maren Bennewitz

**Abstract**—For scene understanding in unstructured environments, an accurate and uncertainty-aware metric-semantic mapping is required to enable informed action selection by autonomous systems. Existing mapping methods often suffer from overconfident semantic predictions, and sparse and noisy depth sensing, leading to inconsistent map representations. In this paper, we therefore introduce EvidMTL, a multi-task learning framework that uses evidential heads for depth estimation and semantic segmentation, enabling uncertainty-aware inference from monocular RGB images. To enable uncertainty-calibrated evidential multi-task learning, we propose a novel evidential depth loss function that jointly optimizes the belief strength of the depth prediction in conjunction with evidential segmentation loss. Building on this, we present EvidKimera, an uncertainty-aware semantic surface mapping framework, which uses evidential depth and semantics prediction for improved 3D metric-semantic consistency. We train and evaluate EvidMTL on the NYUDepthV2 and assess its zero-shot performance on ScanNetV2, demonstrating superior uncertainty estimation compared to conventional approaches while maintaining comparable depth estimation and semantic segmentation. In zero-shot mapping tests on ScanNetV2, EvidKimera outperforms Kimera by 30% in semantic surface mapping accuracy and consistency, highlighting the benefits of uncertainty-aware mapping and underscoring its potential for real-world robotic applications.

## I. INTRODUCTION

A key aspect of robotic systems is semantic scene understanding, as it enables intelligent interaction [1] in applications such as autonomous driving, agriculture, and household robotics. However, for robots to operate reliably in unstructured environments, they should not only recognize objects and surfaces but also quantify the uncertainty in their scene understanding, as wrong predictions can lead to inconsistent world models and therefore unreliable decision-making.

For scene understanding and mapping, traditional frameworks initially focused on purely geometric representations [2], [3], which construct spatial occupancy maps but lack semantic context. More recent semantic Truncated Signed Distance Field (TSDF) mapping methods [4] have enabled dense volumetric representations by propagating 2D semantic labels into 3D space. However, these methods often suffer from overconfident predictions, unreliable depth sensing, and ambiguous 2D-to-3D label fusion, leading to

All authors are with the Humanoid Robots Lab and the Center for Robotics, University of Bonn, Germany. Rohit Menon, Nils Dengler and Maren Bennewitz are additionally with the Lamarr Institute, Bonn, Germany.

This work has partially been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy, EXC-2070 – 390732324 – PhenoRob, by the DFG grant 459376902 – AID4Crops, and by the BMBF within the Robotics Institute Germany, grant No. 16ME0999.

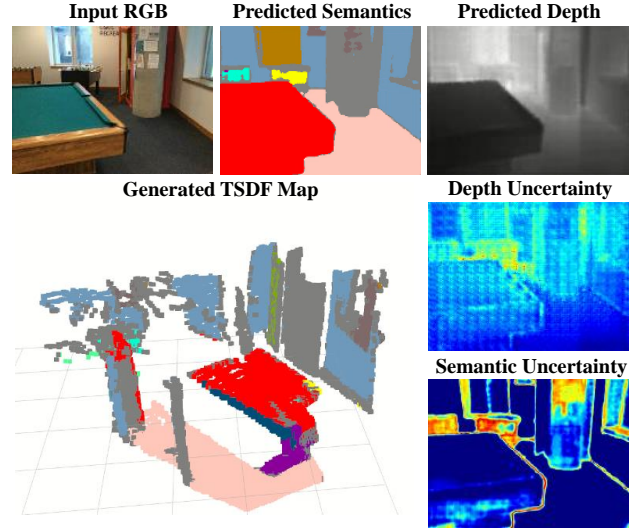


Fig. 1: Visualization of our evidential multi-task perception pipeline. Given RGB image as input on top left, our EvidMTL framework predicts semantic labels and depth (top) along with their corresponding uncertainty estimates (right). The generated TSDF map, shown in bottom left, from our EvidKimera, leverages these uncertainty measurements, only including cells with low depth uncertainty and assigning unknown labels (grey) to regions with high semantics uncertainty.

inconsistent map representations [5]. This motivates the need for uncertainty-aware methods that quantify confidence in both depth and semantics, allowing robots to make more informed decisions. Hence, semantic and depth predictions must not only be more accurate but their uncertainties should also correlate to the actual errors.

Bayesian uncertainty estimation techniques, such as Monte Carlo dropout [6] and ensemble learning [7], have been explored for semantic segmentation but remain computationally expensive due to multiple forward passes or the need for separate network instances [8]. Similarly, while monocular RGB-based depth estimation methods, either standalone [9] or integrated into multi-task frameworks such as SwinMTL [10], help mitigate sparse and noisy depth sensing, they still suffer from overconfidence and unreliability in challenging conditions [11].

To address these limitations, we propose EvidMTL, an evidential multi-task learning framework that extends SwinMTL [10] with uncertainty-aware depth and semantic segmentation. We propose a novel Evidential Scale-Invariant Log (EvidSiLog) loss, which integrates predictive uncertainty regularization with a novel prior-anchored Kullback–Leibler (KL) divergence loss. This KL divergence loss

optimizes the hyperparameters of evidential depth prediction by anchoring them to the noise added ground-truth depth prior, with the predictive uncertainty regularized on the added noise, ensuring stable and effective joint learning of both tasks. Next, we introduce EvidKimera, a semantic TSDF surface mapping framework that extends the multi-view fusion of Kimera [4] by integrating evidential predictions for depth and semantics. EvidKimera employs a weighting strategy for 2D-to-3D label transfer, discounting unreliable depth estimates to mitigate erroneous updates and incorporating viewpoint similarity to prevent the reinforcement of systematic errors.

To demonstrate the benefits of our loss design, we train and validate our networks on the NYUDepthV2 [12] dataset and demonstrate its in-distribution performance. Zero-shot testing on ScanNetV2 [13] shows that EvidMTL achieves superior uncertainty estimation compared to conventional approaches while maintaining comparable depth estimation and semantic segmentation performance. To further explore the impact of uncertainty-aware semantic surface mapping, we conduct zero-shot mapping tests on ScanNetV2, confirming that EvidKimera outperforms Kimera in semantic surface mapping accuracy and consistency. The code of our complete pipeline is available upon publication at [github.com/HumanoidsBonn/evidential\\_mapping](https://github.com/HumanoidsBonn/evidential_mapping).

## II. RELATED WORK

### A. Joint Prediction of Semantic and Depth Information

Dense semantic segmentation is well-established with convolutional architectures like U-Net [14] and DeepLab [15], while monocular depth estimation, pioneered by Eigen *et al.* [16] and later works such as Monodepth [17], predicts dense depth maps from RGB images to address sparse sensing. Jointly learning these modalities reduces computational costs and improves robustness through inter-dependency.

Recent advances favor transformer-based architectures (e.g., Swin Transformer [18] for segmentation, and AdaBins [19] for depth), enhancing global context but increasing complexity. Multi-task learning (MTL) leverages shared representations to improve both tasks, as shown by SwinMTL [10], which uses a shared transformer encoder with task-specific heads for efficiency.

Classical models lack uncertainty quantification, critical for safety-sensitive applications. Bayesian methods like Monte Carlo dropout [6] and deep ensembles [7] provide uncertainty estimates but require multiple forward passes or multiple models, limiting practicality. To overcome these challenges, evidential approaches have been proposed to estimate uncertainty in a single pass for classification [20] and regression tasks [21] respectively.

While Kendall *et al.* [22] addressed multi-task loss balancing via homoscedastic uncertainty, recent work such as EMUFormer [23] employed a student-teacher distillation strategy to predict total uncertainty for joint segmentation and depth estimation. However, it lacks explicit decomposition into epistemic and aleatoric components and relies on a pretrained ensemble teacher. To our knowledge, no

prior work incorporates evidential uncertainty into multi-task learning for this setting. We address this gap by introducing an evidential MTL framework that yields decomposable uncertainties while achieving state-of-the-art performance.

### B. Semantic Mapping

By integrating 3D semantic information derived from 2D images, semantic mapping extends traditional metric maps [2], [3]. Many existing approaches such as Kimera Semantics [4], [24], employ Truncated Signed Distance Fields (TSDFs) for dense volumetric mapping. However, they assign semantic labels using majority voting over hard labels from 2D projections, which limits the reliability of the final 3D map.

An alternative approach incorporates raw segmentation logits to represent class probabilities [25]. While this method improves expressiveness, it still suffers from overconfident predictions and lacks a principled way to estimate uncertainty. To mitigate this issue, Bayesian fusion has been explored [26], but it relies on probabilistic neural networks, which introduce significant computational overhead.

The most relevant works to ours are those of Gan *et al.* [27], Kim *et al.* [28], and Marques *et al.* [29], which all model semantic states using Dirichlet concentration parameters. Gan *et al.* convert one-hot labels from classical segmentation into Dirichlet distributions, ignoring measurement uncertainty and thus lacking a true evidential foundation. Kim *et al.* derive evidence from class probabilities of an evidential network, but discard measurement strength, leading to information loss. Both use costly Bayesian Kernel Inference and infer occupancy from semantic posteriors. Marques *et al.* predict 2D semantic belief states using hard labels and model 2D occupancy in evidential form with additional height maps, without supporting metric depth integration in 3D. In contrast, our method performs occupancy-aware evidential integration that jointly updates semantic and TSDF weight and distance posteriors while preserving uncertainty, enabling robust 3D integration.

To the best of our knowledge, this is the first work to integrate evidential multi-task learning for uncertainty-aware semantic TSDF mapping from monocular images.

## III. OUR APPROACH

To enable reliable 3D scene reconstruction with uncertainty estimation, we propose evidential multi-task learning for depth and semantic predictions from monocular RGB images, along with an uncertainty-aware semantic surface mapping method.

### A. Evidential Multi-Task Learning

For simultaneous semantics and depth prediction, we present **EvidMTL**, an evidential multi-task framework that concurrently predicts semantic labels and depth estimates while explicitly modeling uncertainty, all with a single pass using a shared encoder-decoder architecture, to generate inputs for the proposed mapping approach.

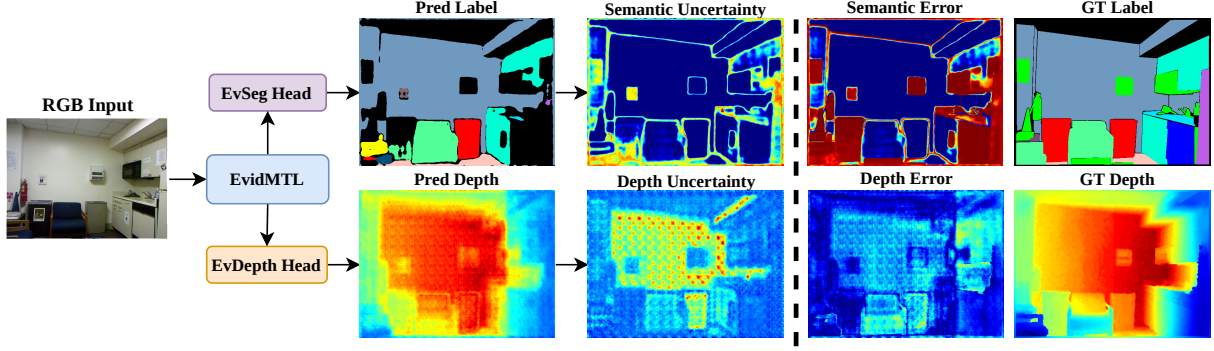


Fig. 2: From an input RGB image our proposed EvidMTL model jointly predicts semantics and depth estimates as well as their uncertainty (left part of the figure). The uncertainty estimates correspond well to the error in the prediction compared to the ground truth (GT) as shown on the right.

Building upon SwinMTL [10], our framework leverages a Swin Transformer [18] for multi-task learning. This hierarchical vision transformer uses shifted windows to enable efficient self-attention, allowing for joint depth estimation and semantic segmentation. In addition, our framework explicitly models evidential uncertainty for both tasks (semantics and depth) and employs a tailored training scheme that mitigates gradient conflicts caused by the evidential regularization losses. Thus, in contrast to SwinMTL, our approach enables stable multi-task learning and robust, uncertainty-aware predictions for subsequent semantic mapping. Fig. 2 shows that from a single RGB input, our EvidMTL model jointly generates not only semantic segmentation and depth prediction but also their uncertainty estimates. The semantic and depth uncertainty correlate with the semantic and depth error respectively. This enables us to generate calibrated uncertainty-aware semantic maps.

### 1) Evidential Depth Prediction

To predict the depth observation from RGB input, we assume a Gaussian distribution [30] and model the predicted depth  $\mu$  with a Gaussian prior, placing its conjugate prior, the Normal Inverse Gamma (NIG) distribution, on the variance  $\sigma^2$ . We replace SwinMTL’s depth prediction head with an evidential regression head [21] to generate the evidential depth parameters from the shared decoder. Thus, for each pixel, instead of predicting only the expected depth  $\mu$ , our evidential depth regression head additionally outputs the hyper-parameters of the NIG distribution  $[\alpha, \beta, \nu]$ , where  $\alpha$  quantifies confidence in the expected depth,  $\beta$  captures uncertainty in the depth noise, and  $\nu$  represents the evidence strength or virtual observation counts for  $\mu$ . The expected depth  $\mathbb{E}[d]$ , expected variance  $\mathbb{E}[\sigma^2]$ , and variance in expected depth  $\text{Var}[d]$  are given as follows [21]:

$$\mathbb{E}[d] = \mu, \quad \mathbb{E}[\sigma^2] = \frac{\beta}{\alpha - 1}, \quad \text{Var}[d] = \frac{\beta}{\nu(\alpha - 1)} \quad (1)$$

Here,  $\mathbb{E}[\sigma^2]$  represents the aleatoric uncertainty in the depth prediction  $u_{al}^d$ , which is irreducible and attributed to the data, while  $\text{Var}[d]$  represents the epistemic uncertainty  $u_{ep}^d$ , which reflects model uncertainty. The total modeled variance or predictive uncertainty is given by  $\sigma_t = \mathbb{E}[\sigma^2] + \text{Var}[d]$ .

We extend the SwinMTL framework by modifying its depth loss to incorporate evidential hyper-parameters. Our novel evidential depth loss  $\mathbb{L}_{ed}$  is defined as:

$$\begin{aligned} \mathbb{L}_{silog} &= \sqrt{\mathbb{E}[(\log d_{gt} - \log \mu)^2] - \lambda \mathbb{E}[\log d_{gt} - \log \mu]^2} \\ \mathbb{L}_{unc} &= \mathbb{E}[\log(\sigma_t^2) - \log((d_{gt} - \mu)^2)] \\ \mathbb{L}_{reg} &= \mathbb{D}_{KL}(\text{NIG}_{pred} \parallel \text{NIG}_{prior}) \\ \mathbb{L}_{ed} &= \mathbb{L}_{silog} + \min(1.0, (\frac{n_{cur}^{ep}}{k \cdot n_{tot}^{ep}})^2)(\lambda_1 \cdot \mathbb{L}_{unc} + \lambda_2 \cdot \mathbb{L}_{reg}) \end{aligned} \quad (2)$$

Here,  $\mathbb{L}_{silog}$  is the Scale-Invariant Log (SiLog) loss [10],  $\mathbb{L}_{unc}$  regularizes predictive uncertainty, and  $\mathbb{L}_{reg}$  is the Kullback–Leibler divergence loss [21] between predicted hyper parameters  $\text{NIG}_{pred}$  and the prior NIG parameters  $\text{NIG}_{prior}$ . Additionally,  $\lambda_1$ ,  $\lambda_2$ , and  $k$  are scaling coefficients,  $n_{cur}^{ep}$  is the current epoch, and  $n_{tot}^{ep}$  is the total number of epochs. To ensure stable learning, we introduce a square-law annealing for  $\mathbb{L}_{reg}$  and  $\mathbb{L}_{unc}$ , gradually increasing its influence during training. This prevents excessive regularization in early epochs while improving uncertainty-aware depth estimation. Additionally,  $\mathbb{L}_{unc}$  enhances robustness by accounting for predictive uncertainty.

### 2) Evidential Semantic Segmentation

In order to account for relative class probabilities and uncertainty in predictions, we model semantic segmentation, a multinomial classification task, as a Dirichlet distribution for evidential prediction. Therefore, we extend SwinMTL’s [10] semantic segmentation head with an *evidence layer* that transforms logits into class-specific evidence values using a softplus activation:

$$e_i = \text{softplus}(z_i), \quad c_i = e_i + 1 \quad (3)$$

where  $z_i$  is the logit for class  $i$ , and  $c_i$  parametrizes the Dirichlet distribution. The expected class probabilities and epistemic uncertainty are computed as:

$$S = \sum c_i, \quad p_i = \frac{c_i}{S}, \quad u_{ep}^s = \frac{K}{S} \quad (4)$$

where  $K$  is the number of semantic classes,  $S$  is the total evidence, and  $u_{ep}^s$  represents epistemic uncertainty, derived from Dempster-Shafer theory [31]. The total evidential se-

semantic segmentation loss is defined as:

$$\begin{aligned}\mathbb{L}_{ece} &= \sum_k^K l_k \cdot (\log S - \log c_k), \quad \mathbb{L}_{KL} = \mathbb{D}_{KL}(\text{Dir}(\mathbf{c}) \parallel \text{Dir}(\mathbf{1})) \\ \mathbb{L}_{es} &= \mathbb{L}_{ece} + \lambda_3 \cdot \min\left(1.0, \frac{n_{cur}^{ep}}{k \cdot n_{tot}^{ep}}\right) \cdot \mathbb{L}_{KL}\end{aligned}\quad (5)$$

with  $\mathbb{L}_{ece}$  as the evidential cross-entropy loss [20],  $\text{Dir}(\mathbf{1})$  the uniform Dirichlet distribution,  $\text{Dir}(\mathbf{c})$  the predicted distribution, and  $\mathbb{D}_{KL}$  the Kullback–Leibler divergence.  $\mathbb{L}_{KL}$  acts as a regularizer, mitigating overconfident predictions. This formulation ensures stable multi-task training by dynamically adjusting the strength of the regularization terms. In particular, we apply evidential uncertainty modeling to SwinMTL’s semantic segmentation pipeline and employ linear annealing for semantic regularization to prevent conflicts with the squared-law depth regularization.

### B. Evidential Semantic Surface Mapping

Fig. 3 shows an overview of our proposed architecture and its three components. It comprises three modules: (1) our EvidMTL network predicting depth and semantic segmentation with uncertainty, (2) a cloud creator fusing predictions into an evidential semantic point cloud, and (3) a mapping framework refining the global metric-semantic map via multi-view uncertainty-weighted integration. The individual components are described in the following.

To combine the output of our evidential multi-task network into a meaningful map representation, we propose an uncertainty-aware semantic TSDF mapping framework that integrates an evidential semantic point cloud, formed by fusing the depth and semantic predictions:

$$\mathcal{P} = \left\{ p_i = (\mathbf{x}_i, rgb, u_{ep}^d, u_{al}^d, c_{i1}, \dots, c_{iK}) \mid i = 1, \dots, N \right\} \quad (6)$$

where  $\mathbf{x}_i = (x_i, y_i, z_i)$  denotes 3D coordinates,  $rgb$  represents color, and  $z_i$  is the expected depth  $\mu$  in the camera frame.

#### 1) Evidential Depth Integration

In comparison to traditional TSDF mapping frameworks such as Voxblox [3] and KinectFusion [32], that assign TSDF weights based on an inverse square law of the depth distance, we propose to incorporate uncertainty-aware weighting. Specifically, we compute the total uncertainty and update the TSDF weights as:

$$u_{tot}^d = u_{ep}^d + u_{al}^d, \quad w_m = \frac{1}{u_{tot}^d}, \quad w_{post} = w_{prior} + w_m \quad (7)$$

where  $u_{tot}^d$  represents the total depth uncertainty,  $w_m$  is the measurement weight, and  $w_{post}$  is the updated weight after incorporating the prior information.

In addition to updating TSDF weights, we also maintain a separate voxel-wise epistemic uncertainty. This uncertainty is updated in a Bayesian manner, using the harmonic mean of the prior and measurement epistemic uncertainties:

$$\frac{1}{u_{ep}^{post}} = \frac{1}{u_{ep}^{prior}} + \frac{1}{u_{ep}^m} \quad (8)$$

This formulation ensures that the epistemic uncertainty is

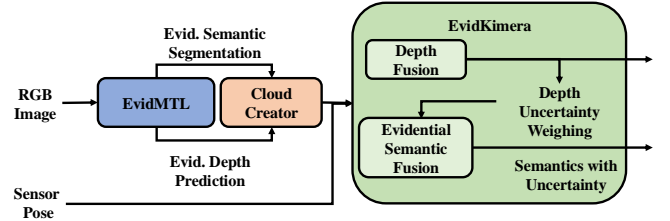


Fig. 3: Our semantic evidential mapping framework: The RGB image is processed through EvidMTL (left), our evidential multi-task depth-semantic segmentation network. The resulting semantic point cloud undergoes uncertainty-weighted Bayesian fusion for the TSDF layer, whereas the evidential semantic predictions are used as the measurements for updating the voxel semantic priors in EvidKimera (right). The final uncertainty-weighted integration refines the semantic voxel posteriors. The mapping framework outputs metric-semantic information with corresponding uncertainties.

refined progressively as more observations are incorporated. Unlike conventional TSDF frameworks that solely rely on depth confidence heuristics, our approach explicitly accounts for both aleatoric and epistemic uncertainties, leading to a more robust and uncertainty-aware reconstruction.

#### 2) Evidential Semantic Integration

Conventional semantic mapping uses majority voting [33] or Bayesian fusion [26], converting between probabilities and distributions. Instead, we represent each voxel’s semantic state as a Dirichlet distribution, exploiting its conjugate-prior property for end-to-end fusion of prior, measurement, and posterior without extra conversions [34]. This preserves uncertainty and accumulates evidence consistently.

We extend the  $K$  semantic classes with two additional states, free space ( $F$ ) and unknown ( $U$ ), by augmenting the Dirichlet concentration vector to length  $L = K + 2$ . Differentiating the 2D background class from the voxel-level unknown class allows us to distinguish confident unknowns from genuine uncertainty. Given a new measurement  $\text{Dir}(\mathbf{c}^m)$ , the posterior update follows:

$$\mathbf{c}^{post} = \mathbf{c}^{prior} + \lambda \mathbf{e}^m, \quad (9)$$

where  $\lambda$  controls the influence of the measurement relative to prior evidence, and  $\mathbf{e}^m = (\mathbf{c}^m - \mathbf{1})$  represents the evidential belief of the measurement. The class probabilities and hard label assignment are computed as:

$$p_k = \frac{c_k}{S}, \quad S = \sum_{k=1}^L c_k. \quad (10)$$

$$\hat{k} = \begin{cases} \arg \max_k c_k, & \text{if } u_{ep}^s < \tau \\ U, & \text{otherwise.} \end{cases} \quad (11)$$

where  $\tau$  is the uncertainty threshold factor. This formulation ensures a recursive evidential update, maintaining probabilistic structure while accumulating confidence from multiple observations.

#### 3) Depth Measurement Weighting

Reliable depth estimates are critical for accurate 2D-to-3D semantic fusion. Instead of assuming uniform confidence,

we introduce an *occupancy belief weighting* mechanism that conditions semantic updates on depth uncertainty, ensuring geometric consistency.

**Occupancy belief modeling:** Unlike prior approaches that treat geometric and semantic uncertainties separately, we explicitly incorporate TSDF-based occupancy confidence into semantic updates. The occupancy belief  $o$  is defined as:

$$o = w^m \cdot \begin{cases} 1 - \exp(-|d^m|), & \text{if } d^m > 0 \text{ (free)} \\ \exp(-|d^m|), & \text{if } d^m < 0 \text{ (occupied)} \end{cases} \quad (12)$$

where  $w^m$  is the TSDF weight, and  $d^m$  is the signed distance function (SDF) value. The depth weighting factor  $\lambda_d$  is then computed as:

$$\lambda_d = \max(o, \epsilon) \quad (13)$$

where  $\epsilon = 0.01$ , ensuring that low-confidence measurements contribute minimally.

**Handling free space and unknown regions:** To maintain consistency, semantic updates incorporate depth-aware adjustments:

$$\mathbf{e}^m = \begin{cases} \lambda_d, & d > 0 \text{ (free space)} \\ \mathbf{e}^m, & d \leq 0 \text{ (occupied space)} \\ \lambda_d, & w < \epsilon \text{ (unknown)}. \end{cases} \quad (14)$$

By explicitly modeling free-space evidence and preserving uncertainty in unknown regions, this approach ensures robustness to measurement noise while preventing erroneous semantic updates.

#### 4) Viewpoint Similarity Discounting

To prevent systematic error accumulation, we weight semantic observations by a viewpoint-dissimilarity factor [35], so that measurements from similar views are discounted.

By integrating these two components into the weighting factor, our framework ensures a robust and uncertainty-aware semantic fusion that maintains consistency across multiple observations while preventing erroneous updates.

### IV. EXPERIMENTAL EVALUATION

The experiments are designed to demonstrate that: (1) The proposed combination of SiLog,  $\text{KL}_\mu$ , and uncertainty regularization loss functions in our EvidMTL framework facilitate the generation of calibrated uncertainty estimates. (2) Our EvidMTL network achieves comparable depth and semantic prediction while delivering superior uncertainty estimation compared to a state-of-the-art baseline, particularly on out-of-distribution data. (3) Our evidential mapping framework, EvidKimera, in conjunction with the outputs of our EvidMTL network, produces an accurate and uncertainty-aware semantic map compared to conventional mapping.

#### A. EvidMTL Evaluation

##### 1) Metrics and Baseline Methods

To evaluate our proposed *EvidMTL* model, we report the following metrics:

- **mIoU:** Mean Intersection-over-Union for semantic segmentation.

- **Pixel Acc:** Pixel accuracy for semantic segmentation.
- **Seg ECE:** Expected Calibration Error; measures alignment between predicted semantic uncertainty (Eq. (4)) and actual segmentation errors.
- **RMSE:** Root Mean Squared Error for depth prediction.
- **Depth NLL:** Negative Log-Likelihood; penalizes inaccuracy or overconfident depth predictions.
- **Normalized Depth ECE:** Expected Calibration Error; measures alignment between predicted depth uncertainty (Eq. (1)) and actual depth errors normalized by the maximum depth of 10m.
- $\nu$ : Strength of predicted depth evidence.

#### 2) Training and Evaluation Setup

We train and validate our networks on the NYUDepthV2 dataset [12], which comprises 795 training images and 694 validation images. All models were then trained on single NVIDIA RTX A6000 GPU with identical hyper-parameters: batch size 4, 500 epochs, learning rate linearly warmed from  $2 \times 10^{-5}$  to  $2 \times 10^{-4}$ , layer-wise decay factor of 0.9, and weight decay of 0.05. Out-of-distribution performance was assessed on the ScanNetV2 dataset [13]. To harmonize semantic annotations, NYU40 labels were remapped to the ScanNet22 ontology, an extension of ScanNet20 with two additional classes for “other furniture” and “other structures”. For a fair comparison, we re-implemented and retrained the SwinMTL baseline [10] on these ScanNet22 labels, employing the Swin V2 Base SMIM backbone [36], [37].

#### 3) Evaluation Results on NYUDepthV2 Dataset

We keep the segmentation head and its cross-entropy loss unchanged and vary only the depth branch. Two main reconstruction terms are considered: the negative log-likelihood (**NLL**) of Amini *et al.* [21] and the scale-invariant logarithmic loss (**SiLog**, Eq. 2). Each can be augmented by one of two evidence regularisers: (i) the original Amini penalty (**Reg**), and (ii) a KL divergence to a conjugate prior with three flavours. **KL** uses *predicted* mean and  $\beta$  together with weak hyper-parameters ( $\alpha=1.01, \nu=0.001$ ); **KL <sub>$\mu$</sub>**  adopts the *ground-truth* depth as prior mean, a fixed  $\beta=0.1$ , and strong hyper-parameters ( $\alpha=2.0, \nu=1.0$ ); **KL <sub>$\mu_n$</sub>**  employs the ground-truth depth perturbed by Gaussian noise and recalculates  $\beta$  from the expected variance, again with weak hyper-parameters. Adding the predictive-uncertainty penalty  $\mathcal{L}_{\text{unc}}$  to SiLog is denoted **EvidSiLog**. Combining {NLL, SiLog, EvidSiLog} with {Reg, KL, KL <sub>$\mu$</sub> , KL <sub>$\mu_n$</sub> } yields eight multi-task configurations; three single-task ablations freeze one of the heads. All models are trained on NYUDepthV2 with only 795 images, making uncertainty learning especially challenging.

The evaluation results are shown in Table I. NLL-based losses struggle with the small data regime, producing the worst depth accuracy (RMSE up to 3.00m) and calibration (Depth-ECE 0.21) even after regularisation. Switching to SiLog reduces RMSE to  $\approx 0.48\text{m}$ , yet naïve regularisers either drive the model into extreme under-confidence (SiLog+Reg,  $\nu \sim 10^{-6}$ , Depth-ECE 2.08) or over-confidence



Type	Model	mIOU $\uparrow$	Pixel Acc $\uparrow$	Seg. ECE $\downarrow$	RMSE $\downarrow$	Depth NLL $\downarrow$	Normalized Depth ECE $\downarrow$	$\nu$
SE	EvidSeg	0.49	0.75	0.09	—	—	—	—
SE	EvidDepth (EvidSiLog+KL $_{\mu_n}$ )	—	—	—	0.48	0.70	0.03	1.23
SE	EvidDepth (NLL+Reg [21])	—	—	—	1.27	1.41	0.21	0.16
MN	SwinMTL [10]	0.48	<b>0.77</b>	—	<b>0.46</b>	—	—	—
ME	NLL+Reg [21]	0.51	0.76	0.10	3.00	2.65	0.21	$2.0 \times 10^{-3}$
ME	NLL+KL [21]	0.52	0.75	0.09	1.23	1.37	0.05	3.024
ME	SiLog+Reg	<b>0.53</b>	0.77	0.08	0.47	3.96	2.08	$1.0 \times 10^{-6}$
ME	EvidSiLog+Reg	0.52	0.76	0.09	0.48	2.43	0.19	$1.0 \times 10^{-3}$
ME	SiLog+KL	0.53	0.76	0.09	0.48	3.05	0.81	31.61
ME	SiLog+KL $_{\mu_n}$	0.53	0.76	0.09	0.48	1.57	0.09	31.91
ME	EvidSiLog+KL $_{\mu}$ (Ours)	0.51	0.76	0.09	0.48	0.86	0.05	293
ME	EvidSiLog+KL $_{\mu_n}$ (Our EvidMTL)	0.52	0.76	0.08	0.48	<b>0.68</b>	<b>0.03</b>	1.22

TABLE I: Semantic and depth performance of the baseline **SwinMTL** and our network variants on the NYUDepthV2 validation split. **S**=single-task, **M** = multi-task; **E**=evidential, **N**=normal (non-evidential). We evaluate eight distinct combinations of depth reconstruction terms (NLL, SiLog, EvidSiLog) and evidence regularizers (Reg, KL, KL $_{\mu}$ , KL $_{\mu_n}$ ). The evidence strength  $\nu$  indicates how confidently the model explains aleatoric versus epistemic uncertainty. Key observations: (1) *SiLog* consistently lowers RMSE relative to NLL. (2) Only when SiLog loss is paired with the uncertainty loss and KL with priors, the models learns to adjust  $\nu$  around the prior, yielding the best Depth-ECE. (3) Our final model, **EvidMTL** (EvidSiLog+KL $_{\mu_n}$ ), matches SwinMTL’s segmentation and depth accuracy while providing well-calibrated depth uncertainty for no extra network cost.

Model	mIOU $\uparrow$	Pixel Acc $\uparrow$	Seg. ECE $\downarrow$	RMSE $\downarrow$	Depth NLL $\downarrow$	Normalized Depth ECE $\downarrow$
SwinMTL [10]	0.35	0.57	—	<b>0.42</b>	—	—
SiLog+Reg	0.35	0.59	<b>0.04</b>	0.43	1.09	0.28
SiLog+KL	0.36	0.59	0.04	0.43	0.96	0.22
SiLog+KL $_{\mu_n}$	0.35	0.59	0.04	0.44	0.78	0.09
EvidSiLog+KL $_{\mu}$ (Ours)	0.35	0.58	0.04	0.43	1.73	0.05
EvidSiLog+KL $_{\mu_n}$ (Our EvidMTL)	<b>0.36</b>	0.59	0.04	0.42	<b>0.72</b>	<b>0.04</b>

TABLE II: Zero-shot semantic and depth performance on **ScanNetV2** (out-of-distribution w.r.t. NYUDepthV2 training). Bold numbers mark the best result in each column. Evidential losses with the noisy-prior KL $_{\mu_n}$  term (last row) deliver the lowest Depth-NLL and Depth-ECE while retaining SwinMTL-level segmentation and depth accuracy.

(SiLog+KL,  $\nu = 31.6$ , Depth-ECE 0.81). Imposing a *noisy* ground-truth prior (SiLog+KL $_{\mu_n}$ ) moderates evidence strength and cuts NLL from 3.05 to 1.57, but reliability improves decisively only when SiLog is paired with  $\mathcal{L}_{unc}$ . The resulting **EvidSiLog+KL $_{\mu}$**  and **EvidSiLog+KL $_{\mu_n}$**  delivers the best calibrated depth uncertainty predictions while being comparable in mIOU and RMSE among multi-task models. Although both models have similar total predictive uncertainty calibration, EvidSiLog+KL $_{\mu}$  is an over-confident model with high evidence strength whereas EvidSiLog+KL $_{\mu_n}$  is less confident in its depth prediction.

The single-task EvidDepth (EvidSiLog+KL $_{\mu_n}$ ) shows no measurable benefit over the multi-task counterpart, its NLL (0.70) and Depth-ECE (0.03) are effectively identical to the multi-task values (0.68 / 0.03). By retaining the same depth performance while adding only a lightweight segmentation head, the multi-task configuration provides calibrated predictions for *both* tasks at no additional cost.

#### 4) Out-of-Distribution Testing on ScanNetV2 Dataset

To assess zero-shot generalization, we evaluate **SwinMTL**, **SiLog+Reg**, **SiLog+KL**, **SiLog+KL $_{\mu}$ +Unc**, **EvidSiLog+KL $_{\mu}$** , and our full **EvidSiLog+KL $_{\mu_n}$**  on ten randomly selected scenes from the ScanNetV2 dataset [13]. Since all models are trained exclusively on NYUDepthV2, ScanNetV2 introduces a significant domain shift while still containing indoor scenes.

The results in Table II show that: (1) Our proposed EvidSiLog+KL $_{\mu_n}$  exceeds SwinMTL in segmentation accuracy (mIOU= 0.36, Pixel Acc = 0.59) and matches its lowest

depth RMSE (0.42 m) while yielding the lowest Depth NLL (0.72) and Depth ECE (0.04). (2) Incorporating the uncertainty loss with a noisy ground-truth prior (KL $_{\mu_n}$ ) consistently calibrates depth uncertainty better than either the Reg or plain KL terms (Depth ECE drops from 0.28/0.22 to 0.04). (3) Although SiLog+Reg and SiLog+KL retain competitive mIOU and RMSE, their higher NLL and ECE indicate unreliable confidence estimates, potentially problematic for downstream mapping. Qualitative error-versus-uncertainty visualizations in Fig. 4 corroborate that EvidMTL produces the most plausible depth-uncertainty maps under domain shift. All models have an average inference time less than 100ms on an RTX3080Ti GPU. In the next section, we explore the importance of introducing uncertainty for the network in the mapping task.

### B. Evidential Semantic Mapping Evaluation

#### 1) Metrics and Baseline Methods

To evaluate our evidential semantic mapping framework, we use the following metrics to assess the quality of the generated semantic TSDF maps:

- **3D mIoU**: Mean cumulative Intersection-over-Union across all scenes for voxels with ground truth correspondence within a threshold (equal to voxel size).
- **Segmentation Voxel Accuracy**: Fraction of voxels with correct semantic labels.
- **Segmentation Voxel ECE**: Expected Calibration Error; measures the correlation between predicted voxel uncertainty and semantic label error.

Framework	MTL Model	Voxel Repr.	Seg. Repr.	TSDF Weight	3D mIoU $\uparrow$	Seg. Voxel Acc $\uparrow$	Seg. Voxel ECE $\downarrow$
Kimera	2DGT (GT labels)	hard	hard	$1/d^2$	<b>0.54</b>	<b>0.69</b>	–
	SwinMTL [10]	hard	hard	$1/d^2$	0.22	0.36	–
EvidKimera (Ours)	SwinMTL ( $u_{ep}^s < 0.5$ )	evid	logits $\rightarrow$ evid	$1/d^2$	0.02	0.05	0.21
	SwinMTL ( $u_{ep}^s < 0.5$ )	evid	one-hot $\rightarrow$ evid	$1/d^2$	0.01	0.02	–
	EvidMTL ( $u_{ep}^s < 0.3$ )	evid	evid	$1/d^2$	0.24	0.36	0.43
	EvidMTL ( $u_{ep}^s < 0.3$ )	evid	evid	$1/u_{tot}^d$	0.26	0.39	<b>0.43</b>
	EvidMTL ( $u_{ep}^s < 0.5$ )	evid	evid	$1/u_{tot}^d$	0.29	0.43	0.43

TABLE III: Zero-shot 3D semantic surface mapping evaluation on ScanNetV2. Columns denote mapping framework, multi-task learning (MTL) model, semantic representation used for voxels and pixels (hard, evidential), TSDF weighting (distance-based  $1/d^2$  or uncertainty-based  $1/u_{tot}^d$ ), and the evaluation metrics.

We compare the following mapping variants, as summarized in Table III:

- **Kimera + 2DGT**: Hard voxel and hard segmentation using ground-truth labels; TSDF weight  $1/d^2$ .
- **Kimera + SwinMTL**: Hard voxel and hard segmentation via majority vote of SwinMTL predictions; TSDF weight  $1/d^2$ .
- **EvidKimera + SwinMTL-Logits**: Evidential fusion of SwinMTL logits (evid/logits $\rightarrow$  evid); TSDF weight  $1/d^2$ .
- **EvidKimera + SwinMTL-OneHot**: Evidential fusion of SwinMTL hard labels encoded as evidence (evid/one hot $\rightarrow$  evid); TSDF weight  $1/d^2$ .
- **EvidKimera + EvidMTL**: Full EvidMTL Dirichlet evidential fusion (evid/evid); TSDF weight  $1/d^2$ .
- **EvidKimera + EvidMTL** ( $u_{ep}^s < 0.3$ ): Full EvidMTL Dirichlet evidential fusion with total-depth-uncertainty TSDF weighting ( $1/u_{tot}^d$ ) and segmentation-uncertainty threshold 0.3.
- **EvidKimera + EvidMTL** ( $u_{ep}^s < 0.5$ ): Same as above with segmentation-uncertainty threshold 0.5.

All experiments run online on Ubuntu 20.04 with ROS Noetic on an 11th-Gen Intel i9-11900K CPU and NVIDIA RTX 3080 Ti GPU, with EvidKimera running at 1 Hz. To mimic real application scenarios, we perform zero-shot inference without fine-tuning the networks.

## 2) Zero-shot Evaluation Results

Our zero-shot mapping experiments (see Tab. III) show that even an idealized pipeline using perfect 2D semantics and depth (“Kimera + 2DGT”) achieves only 0.54 3D mIoU and 69% Segmentation Voxel Accuracy. We attribute this gap to inevitable 2D-to-3D projection errors arising from view-to-view inconsistencies and to the overconfident nature of hard labels, which cannot express uncertainty about occlusions or noisy depth. When replacing the ground-truth semantics with the SwinMTL predictions without any uncertainty modelling (“Kimera + SwinMTL”), performance collapses (3D mIoU = 0.22, accuracy = 0.36), underscoring how overconfident but incorrect 2D predictions further degrade the 3D reconstruction.

Interpreting SwinMTL outputs as evidence, either by casting logits into a Dirichlet form or by converting hard one-hot labels into pseudo-evidence, yields no semantic recovery (3D mIoU = 0.02, accuracy = 0.04), despite reasonable calibration error (ECE = 0.21). This confirms that naive evidential encoding of 2D outputs does not compensate for model

uncertainty. In contrast, our full EvidMTL Dirichlet fusion quantifies epistemic uncertainty at both pixel and voxel levels, recovering much of the lost performance (3D mIoU = 0.24, accuracy = 0.36) under distance-based TSDF weighting. By further adopting uncertainty-aware TSDF weights ( $1/u_{tot}^d$ ), we improve the results to 29% mIoU and 43% accuracy while maintaining a Seg. Voxel ECE of = 0.43 which is on par with state-of-the-art 2D calibration techniques [38]. These findings demonstrate that principled evidential fusion, combined with uncertainty-aware integration, substantially closes the gap to the ground-truth upper bound and provides reliable confidence estimates for downstream tasks.

## V. CONCLUSION

In this work, we introduce EvidMTL and EvidKimera, which integrate an evidential multi-task learning framework for uncertainty-aware semantic surface mapping from monocular images. Our EvidMTL network, featuring two novel loss terms, jointly predicts semantic segmentation and depth estimation while explicitly modeling uncertainty, thereby enhancing prediction reliability and consistency. Furthermore, our evidential semantic mapping framework EvidKimera leverages uncertainty quantification in both semantic and depth predictions to generate an uncertainty-aware semantic TSDF map. Compared to baselines, EvidMTL achieves comparable performance in depth and semantic prediction while providing superior uncertainty estimation, particularly in depth uncertainty estimation, which in turn boosts 3D mapping performance EvidKimera framework. To our knowledge, this is the first evidential multi-task learning framework for semantic TSDF mapping.

## REFERENCES

- [1] S. Garg, N. Sünderhauf, F. Dayoub, D. Morrison, A. Cosgun, G. Carneiro, Q. Wu, T.-J. Chin, I. Reid, S. Gould, *et al.*, “Semantics for robotic mapping, perception and interaction: A survey,” *Foundations and Trends® in Robotics*, 2020.
- [2] A. Hornung, K. M. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard, “Octomap: An efficient probabilistic 3d mapping framework based on octrees,” *Autonomous Robots*, 2013.
- [3] H. Oleynikova, Z. Taylor, M. Fehr, R. Siegwart, and J. Nieto, “Voxblox: Incremental 3d euclidean signed distance fields for on-board mav planning,” in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2017.
- [4] A. Rosinol, M. Abate, Y. Chang, and L. Carlone, “Kimera: An open-source library for real-time metric-semantic localization and mapping,” in *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2020.

- [5] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in *Proc. of the Intl. Conf. on Machine Learning*, ser. Proceedings of Machine Learning Research. PMLR, 2017.
- [6] Y. Gal and Z. Ghahramani, “Dropout as a bayesian approximation: Representing model uncertainty in deep learning,” in *Proc. of the Intl. Conf. on Machine Learning*. PMLR, 2016.
- [7] B. Lakshminarayanan, A. Pritzel, and C. Blundell, “Simple and scalable predictive uncertainty estimation using deep ensembles,” *Advances in Neural Information Processing Systems*, 2017.
- [8] M. Sharma, S. Farquhar, E. Nalisnick, and T. Rainforth, “Do bayesian neural networks need to be fully stochastic?” in *Proc. of the Intl. Conf. on Artificial Intelligence and Statistics (AIS)*. PMLR, 2023.
- [9] Y. Ming, X. Meng, C. Fan, and H. Yu, “Deep learning for monocular depth estimation: A review,” *Neurocomputing*, 2021.
- [10] P. Taghavi, R. Langari, and G. Pandey, “SwinMTL: A shared architecture for simultaneous depth estimation and semantic segmentation from monocular camera images,” in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2024.
- [11] S. Gasperini, N. Morbitzer, H. Jung, N. Navab, and F. Tombari, “Robust monocular depth estimation under challenging conditions,” in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [12] P. K. Nathan Silberman, Derek Hoiem and R. Fergus, “Indoor segmentation and support inference from rgb-d images,” in *Proc. of the Europ. Conf. on Computer Vision (ECCV)*, 2012.
- [13] A. Dai, M. Nießner, M. Zollöfer, S. Izadi, and C. Theobalt, “Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface re-integration,” *ACM Transactions on Graphics (TOG)*, 2017.
- [14] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference*. Springer, 2015.
- [15] C. Liang-Chieh, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille, “Semantic image segmentation with deep convolutional nets and fully connected crfs,” in *Proc. of the Intl. Conf. on Learning Representations*, 2015.
- [16] D. Eigen, C. Puhrsch, and R. Fergus, “Depth map prediction from a single image using a multi-scale deep network,” *Advances in Neural Information Processing Systems*, 2014.
- [17] C. Godard, O. Mac Aodha, and G. J. Brostow, “Unsupervised monocular depth estimation with left-right consistency,” in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [18] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, et al., “Swin transformer v2: Scaling up capacity and resolution,” in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [19] S. F. Bhat, I. Alhashim, and P. Wonka, “Adabins: Depth estimation using adaptive bins,” in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [20] M. Sensoy, L. Kaplan, and M. Kandemir, “Evidential deep learning to quantify classification uncertainty,” *Advances in Neural Information Processing Systems*, 2018.
- [21] A. Amini, W. Schwarting, A. Soleimany, and D. Rus, “Deep evidential regression,” *Advances in Neural Information Processing Systems*, 2020.
- [22] A. Kendall, Y. Gal, and R. Cipolla, “Multi-task learning using uncertainty to weigh losses for scene geometry and semantics,” in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [23] S. Landgraf, M. Hillemann, T. Kapler, and M. Ulrich, “Efficient multi-task uncertainties for joint semantic segmentation and monocular depth estimation,” in *DAGM German Conference on Pattern Recognition*. Springer, 2024.
- [24] 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 Robotics and Automation Letters (RA-L)*, 2019.
- [25] N. Sünderhauf, T. T. Pham, Y. Latif, M. Milford, and I. Reid, “Meaningful maps with object-oriented semantic mapping,” in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2017.
- [26] D. Morilla-Cabello, L. Mur-Labadia, R. Martinez-Cantin, and E. Montijano, “Robust fusion for bayesian semantic mapping,” in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2023.
- [27] L. Gan, R. Zhang, J. W. Grizzle, R. M. Eustice, and M. Ghaffari, “Bayesian spatial kernel smoothing for scalable dense semantic mapping,” *IEEE Robotics and Automation Letters (RA-L)*, 2020.
- [28] J. Kim, J. Seo, and J. Min, “Evidential semantic mapping in off-road environments with uncertainty-aware bayesian kernel inference,” in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2024.
- [29] J. M. C. Marques, N. Dengler, T. Zaenker, J. Mucke, S. Wang, M. Bennewitz, and K. Hauser, “Map space belief prediction for manipulation-enhanced mapping,” 2025.
- [30] A. Belhedi, A. Bartoli, S. Bourgeois, V. Gay-Bellile, K. Hamrouni, and P. Sayd, “Noise modelling in time-of-flight sensors with application to depth noise removal and uncertainty estimation in three-dimensional measurement,” *IET Computer Vision*, 2015.
- [31] G. Shafer, “Dempster-shafer theory,” *Encyclopedia of artificial intelligence*, 1992.
- [32] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon, “Kinectfusion: Real-time dense surface mapping and tracking,” in *2011 10th IEEE international symposium on mixed and augmented reality*. IEEE, 2011.
- [33] Z. Zhao and X. Chen, “Semantic mapping for object category and structural class,” in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2014.
- [34] B. A. Frigiyk, A. Kapila, and M. R. Gupta, “Introduction to the dirichlet distribution and related processes,” *Department of Electrical Engineering, University of Washington, UWEETR-2010-0006*, 2010.
- [35] R. Menon, T. Zaenker, N. Dengler, and M. Bennewitz, “Nbv-sc: Next best view planning based on shape completion for fruit mapping and reconstruction,” in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2023.
- [36] Z. Xie, Z. Geng, J. Hu, Z. Zhang, H. Hu, and Y. Cao, “Revealing the dark secrets of masked image modeling,” in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [37] D. Kim, W. Ka, P. Ahn, D. Joo, S. Chun, and J. Kim, “Global-local path networks for monocular depth estimation with vertical cutdepth,” arXiv preprint arXiv:2201.07436.
- [38] J. M. C. Marques, A. J. Zhai, S. Wang, and K. Hauser, “On the overconfidence problem in semantic 3d mapping,” in *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2024.

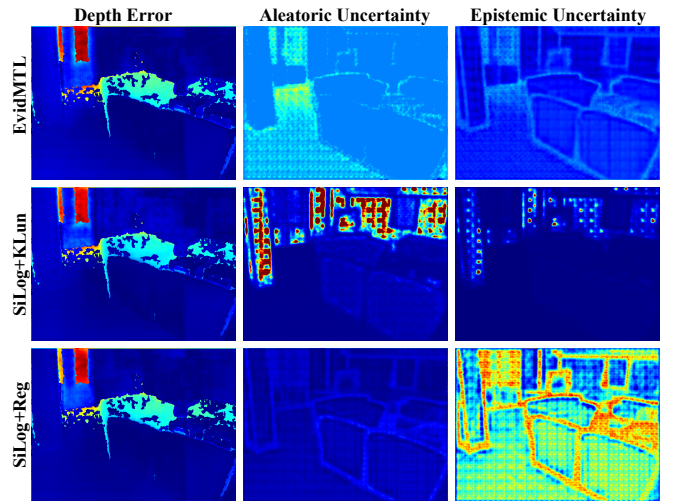


Fig. 4: The columns show the depth error, aleatoric and epistemic uncertainty for a random scene from the ScanNetV2 dataset for the zero-shot evaluation of EvidMTL (EvidSiLog+KL $_{\mu_n}$ ), SiLog+KL $_{\mu_n}$ , and SiLog+Reg. The error and uncertainty increase from blue to red. The high error red spots on the top are windows. SiLog+KL $_{\mu_n}$  (middle row) attributes errors to aleatoric uncertainty whereas SiLog+Reg (bottom row) attributes it to epistemic uncertainty. Our EvidMTL (top row) correctly shows epistemic uncertainty at object boundaries and aleatoric uncertainty on windows and low texture carpets on the floor.