# Analytic Bayesian Uncertainty for LiDAR Segmentation: A Single-pass Generative Approach

Hanieh Shojaei Miandashti[1], Qianqian Zou[2], and Claus Brenner[1]

[1] Institute of Cartography and Geoinformatics, Leibniz University Hannover, Germany
{hanieh.shojaei, claus.brenner}@ikg.uni-hannover.de
[2] Swiss Data Science Center, EPFL, Switzerland
qianqian.zou@epfl.ch

**Abstract.** Accurate predictions with well-calibrated uncertainties are vital for safety-critical LiDAR scene understanding in autonomous driving and robotics. While Bayesian Neural Networks and deep ensembles are effective, they are computationally expensive, requiring multiple forward passes or independently trained models. We propose GMM-NIG, a Bayesian generative classification head that extends the GMMSeg framework by placing conjugate Normal-Inverse-Gamma (NIG) priors over class-conditional GMM parameters to obtain posterior-predictive uncertainty. While the backbone remains deterministic, GMM-NIG enables analytic closed-form posterior updates in a single training run. In inference, GMM-NIG yields a mixture of Student-t distributions, providing well-calibrated predictive uncertainty without extra computational overhead. On SemanticKITTI and nuScenes, GMM-NIG achieves state-of-the-art calibration with an Adaptive Calibration Error of 2.01%, outperforming deep ensembles, Monte Carlo Dropout, evidential models and post-hoc calibration methods. Beyond calibration, it increases segmentation mIoU by nearly 6% and yields the strongest failure detection performance.

**Keywords:** Epistemic uncertainty · Confidence calibration · Segmentation · Generative classification · Gaussian Mixtures · Failure detection

## 1 Introduction

3D scene understanding from LiDAR point clouds is commonly addressed using semantic segmentation models, with range-view representations widely adopted due to their compact, information-rich panoramic encoding of LiDAR data, enabling real-time deployment [15, 33]. However, these models are often trained as point estimators, producing a single categorical distribution over classes, typically via a softmax layer. While these models achieve strong predictive performance, their confidence estimates are often poorly calibrated, leading to overconfident predictions even when they are incorrect. This limitation is particularly critical in safety-sensitive applications, where reliable uncertainty estimation is necessary to detect likely failure regions and prevent erroneous decision-making.

Uncertainty in predictive models is generally categorized into aleatoric uncertainty, arising from intrinsic noise or ambiguity in the observations, such as noisy data or labels [60], and epistemic uncertainty, which reflects uncertainty in model parameters, typically emerging in regions distant from the training distribution or on unseen samples. Bayesian Neural Networks, including variational inference [21], Monte Carlo dropout (MC Dropout) [17], and Laplace approximations [48], provide a principled approach for modeling predictive uncertainty but introduce substantial computational overhead due to additional variational optimization or repeated stochastic forward passes at inference.

To address this, second-order predictive models have been proposed that represent predictions as distributions over class probabilities, enabling direct uncertainty estimation in a single forward pass. Evidential Deep Learning (EDL) [49] follows this direction, estimating parameters of a second-order Dirichlet distribution within a deterministic framework. While EDL enables efficient uncertainty estimation, recent studies have questioned its theoretical consistency as a model of epistemic uncertainty [5, 6, 25, 39, 50]. In particular, these works argue that second-order uncertainty learned solely through empirical risk minimization may fail to recover the true predictive uncertainty and lacks guarantees of consistency with Bayesian posterior inference. Consequently, EDL may behave more like an energy-based model or an out-of-distribution detector rather than a principled estimator of predictive uncertainty [50].

In this work, we introduce GMM-NIG, a sampling-free uncertainty estimation framework for LiDAR semantic segmentation. Our method builds on the generative classification formulation of GMMSeg [34], which replaces the conventional softmax classifier with class-conditional Gaussian mixture model (GMMs) in the feature space. We extend it by modeling uncertainty over the Gaussian component parameters. Specifically, we impose conjugate Normal-Inverse-Gamma (NIG) priors over the component-wise means and variances and analytically marginalize these parameters at inference time. This yields a mixture of Student-t predictive densities, which provides calibrated predictive uncertainty without stochastic sampling, model ensembling, or retraining.

Our main contributions, validated on two LiDAR datasets across both CNN- and Transformer-based backbones and evaluated against established uncertainty estimation baselines, are summarized as follows:

- **Improved segmentation performance.** Our proposed t-distribution posterior predictive improves segmentation accuracy by modeling richer feature distributions beyond discriminative classifiers.
- **Reliable uncertainty estimation.** GMM-NIG produces well-calibrated uncertainty estimates and improves failure detection compared to existing uncertainty-aware baselines.
- **Single-forward uncertainty quantification.** Our method enables sampling free predictive uncertainty using a single deterministic forward pass, achieving the theoretical benefits of Bayesian inference with the computational speed of a standard deterministic network.

– **Backbone-Agnostic Integration.** Our approach can be seamlessly incorporated into CNN and Transformer-based segmentation models without architectural redesign.

## 2 Related Works

### 2.1 Uncertainty Quantification in Semantic Segmentation

In early *deterministic networks*, uncertainty is often approximated using softmax confidence or softmax entropy [23]. However, these measures are often poorly calibrated [22]. To address this issue, some works explicitly model aleatoric uncertainty to calibrate confidence [12, 26, 41], though these methods necessitate modifications to the training procedure.

Alternatively, *post-hoc calibration methods*, such as temperature scaling [22, 35], calibrate model confidence by rescaling output logits using a learned parameter. Dirichlet calibration [29] extends this approach by modeling predictive probabilities with a Dirichlet distribution and Depth-aware scaling [28] rescales according to LiDAR depth. While simple and effective on in-distribution data, these techniques fail to generalize under distribution shift [47] and lack explicit uncertainty estimation.

*Bayesian models* capture epistemic uncertainty by inferring the posterior distribution of model parameters, yet exact inference remains intractable. Variational inference approximations, such as Bayes by Backprop [9], optimize the Evidence Lower Bound to learn a family of tractable distributions over network weights. Similarly, MC Dropout [17] interprets dropout as a variational approximation, estimating epistemic uncertainty by performing multiple stochastic forward passes during inference with dropout activated. However, they often face poor scalability to large models [18] and the computational overhead due to repeated sampling [44].

*Deep Ensembles (DEs)* [31] are widely regarded as the gold standard for uncertainty estimation by aggregating predictions from multiple independently trained models [24, 32]. While empirical studies have consistently demonstrated that DEs outperform MC Dropout [4, 18, 47], the need to retrain the model multiple times leads to a computational cost that scales with both the ensemble sizes and the complexity of the underlying architecture.

While these approaches have been primarily studied in image-based settings, several works have adapted uncertainty estimation techniques to LiDAR semantic segmentation.

### 2.2 LiDAR Semantic Segmentation

Range-view-based semantic segmentation models are commonly implemented using CNN [11, 42, 56, 58, 61]. RangeNet++ [42] introduced an encoder–decoder fully convolutional network, while SalsaNext [13] enhances contextual modeling through dilated convolutions that enlarge receptive fields. FIDNet [61] and

CENet [11] adopt ResNet encoders with lightweight interpolation-based decoders for efficient feature upsampling, and FRNet [59] further enhances segmentation performance through feature refinement modules. Recent transformer-based approaches, including RangeViT [2] and RangeFormer [27], capture both local and long-range contextual dependencies, addressing limitations of fully convolutional architectures.

To incorporate the prediction uncertainty, MC Dropout has been considered in SalsaNext [13] and subsequent studies [14], while DE has also been explored for reliable LiDAR segmentation [16] as well.

Beyond sampling-based methods, evidential frameworks have been proposed for LiDAR perception [1,43,52]. More recent works compare energy-based methods, ensembles, MC Dropout, and evidential learning on bird's-eye view LiDAR segmentation [60], while Calib3D [28] investigates post-hoc calibration approaches for this task.

### 2.3 Generative Classifiers

However, the aforementioned segmentation approaches remain discriminative, which output predictions directly using a softmax and focus solely on optimizing decision boundaries, often leading to overconfident predictions. To overcome this limitation, several works in the image domain have explored generative classifiers that explicitly model latent feature distributions [46], particularly through Gaussian and GMM-based approaches. However, some Gaussian-based classifiers are still trained within discriminative frameworks [37, 46, 51, 54]. For instance, DDU [44] assumes that, once the model is trained discriminatively, the features of each class follow a Gaussian distribution in the latent space, and uses this assumption to estimate predictive uncertainty, although this assumption lacks theoretical guarantees. In contrast, GMMSeg [34] performs semantic segmentation via likelihood-based training by modeling class-conditional features using Gaussian mixtures instead of softmax normalization. However, the model parameters are still estimated deterministically (point estimation), limiting their ability to represent epistemic uncertainty over distribution parameters.

The classification formulation of our work falls within the family of Bayesian generative models, inferring posterior predictive distributions to deliver well-calibrated outputs. However, unlike DE or MC dropout, our approach allows for single-forward inference. This design combines computational efficiency with Bayesian uncertainty estimation. Our proposed GMM-NIG head expects gaussian feature space, for which we adopt GMMSeg [34] to ensure class-conditional GMM in the latent feature space. The key difference is that GMMSeg estimates the GMM parameters as deterministic point estimates, whereas GMM-NIG places conjugate NIG priors over the component parameters and derives a Student-t posterior predictive distribution by analytic marginalization.

## 3 Methodology

We aim to offer an accurate and sampling-free framework for uncertainty estimation in LiDAR scene segmentation. An overview of our approach is illustrated in Fig. 1, with details described below.

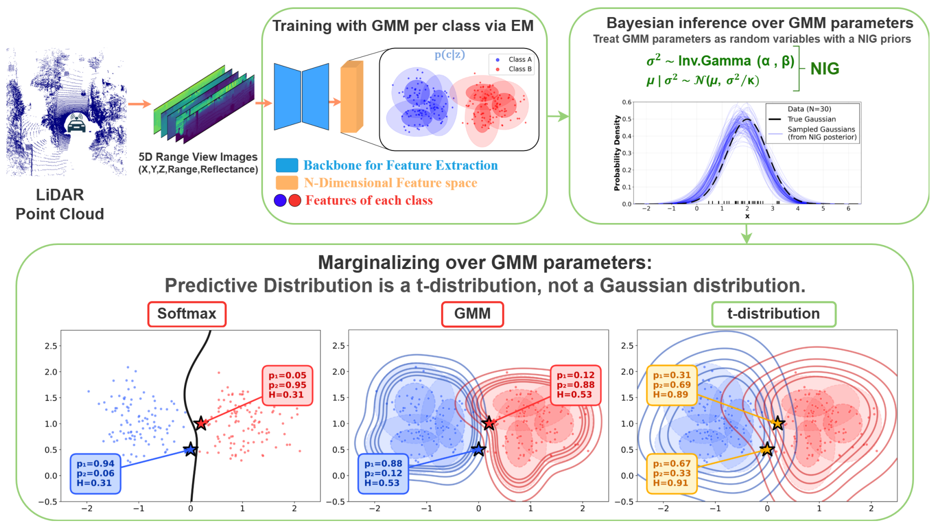



**Fig. 1:** Overview of our approach. For ambiguous points far from the cluster mean, *softmax* and *GMM* produce overconfident probabilities and low *entropy (H)*. In contrast, our NIG-derived heavy-tailed *t-distribution* assigns lower probabilities, yielding higher entropy and better-calibrated uncertainty.

### 3.1 Problem Definition

Let a raw LiDAR point cloud be represented as an unordered set $\mathcal{P}$, where each point contains its Cartesian coordinates $(x, y, z)$ and intensity $i$, and is annotated with a semantic label $c \in \{1, \ldots, C\}$. We transform the unstructured point cloud into a structured representation using spherical projection [42], which projects each 3D point onto a 2D grid of resolution $H \times W$. Each projected pixel is represented by a 5D input feature vector $(r, x, y, z, i)$, where $r = \sqrt{x^2 + y^2 + z^2}$ denotes the range from the LiDAR sensor.

The resulting training data $\mathcal{D} = \{(\mathbf{x}_n, \mathbf{c}_n)\}_{n=1}^{N}$ consists of $N$ pairs, where each input $\mathbf{x}_n \in \mathbb{R}^{H \times W \times 5}$ is a projected range image with five input channels, and $\mathbf{c}_n \in \{1, \ldots, C\}^{H \times W}$ is the corresponding semantic label map. The goal is to learn a model that predicts a semantic label $\hat{c} \in \{1, \ldots, C\}$ for any given pixel with an input vector $\mathbf{x}'$.

To this end, we employ a dense feature extractor $f_\theta : \mathbb{R}^{H \times W \times 5} \to \mathbb{R}^{H \times W \times D}$, typically implemented as an encoder-decoder network, which maps an input image $\mathbf{x}$ to a spatial feature representation $\mathbf{z} = f_\theta(\mathbf{x}) \in \mathbb{R}^{H\times W\times D}$. Since the spatial resolution remains unchanged in semantic segmentation tasks, we simplify notation by omitting $H$ and $W$, and treat each pixel's feature vector as $\mathbf{z} \in \mathbb{R}^D$.

Conventional discriminative classifiers estimate the posterior distribution $p(c|\mathbf{z})$ by first mapping the feature vector $\mathbf{z} \in \mathbb{R}^D$ to a set of class logits in $\mathbb{R}^C$ through a linear projection or $1 \times 1$ convolution, followed by a softmax function to obtain class probabilities. In contrast, we adopt a generative classification approach that models the joint distribution $p(\mathbf{z}, c)$, by learning the class-conditional $p(\mathbf{z}|c)$ and given the class prior $p(c)$, which are commonly assumed to follow a uniform distribution, also adopted in our work. The class posterior is then obtained via Bayes' rule:

$$p(c|\mathbf{z}) = \frac{p(\mathbf{z}|c)\, p(c)}{p(\mathbf{z})} = \frac{p(\mathbf{z}|c)\, p(c)}{\sum_{c'} p(\mathbf{z}|c')\, p(c')}. \tag{1}$$

### 3.2 Generative Classification with GMMs

Inspired by GMMSeg [34], we model the features of each class $c$ using a GMM composed of $M$ multivariate Gaussian components in a $D$-dimensional latent space. To address the practical challenges of directly training generative models discussed in [8], a hybrid approach is adopted, combining the Sinkhorn-EM algorithm [40] for GMM parameter estimation, with discriminative training of a dense feature extractor $f_\theta$. Specifically, the likelihood of a pixel-level feature vector $\mathbf{z} \in \mathbb{R}^D$ for class $c$ is expressed as: $p(\mathbf{z} \mid c; \boldsymbol{\phi}_c) = \sum_{m=1}^{M} \pi_{cm}\, \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_{cm}, \boldsymbol{\Sigma}_{cm})$, where $\pi_{cm}$ denotes the mixing weight of the $m$-th Gaussian component of the class $c$, subject to the simplex constraint $\sum_m \pi_{cm} = 1$. The terms $\boldsymbol{\mu}_{cm} \in \mathbb{R}^D$, $\boldsymbol{\Sigma}_{cm} \in \mathbb{R}^{D\times D}$ represent the mean and covariance, respectively. We denote the complete set of parameters for the GMM of class $c$ as $\boldsymbol{\phi}_c = \{\pi_{cm}, \boldsymbol{\mu}_{cm}, \boldsymbol{\Sigma}_{cm}\}_{m=1}^{M}$.

Sinkhorn-EM algorithm [40] formulates the E-step as an entropy-regularized optimal transport problem solved via Sinkhorn iterations. This replaces point-wise posterior responsibilities with globally coupled, balanced assignments, yielding improved numerical stability and convergence in high-dimensional latent spaces. In this formulation, the mixture weights are assumed to be uniform, i.e., $\pi_{cm} = \frac{1}{M}$, such that class priors are handled exclusively at the semantic prediction stage rather than within the class-conditional GMM.

The Sinkhorn-EM procedure is coupled with discriminative supervision. Motivated by the improved calibration properties of Focal Loss [36], we propose the joint loss function as:

$$\mathcal{L}_{\text{total}} = \lambda_g\, \mathcal{L}_{\text{EM}} + \lambda_d \left(\mathcal{L}_{\text{Lovasz}} + \mathcal{L}_{\text{Focal}}\right), \tag{2}$$

where $\mathcal{L}_{\text{EM}}$ denotes the negative expected log-likelihood of the GMM under the Sinkhorn assignments: $\mathcal{L}_{\text{EM}} = -\sum_{c=1}^{C} \sum_{i:c_i=c} \log \sum_{m=1}^{M} \pi_{cm}\, \mathcal{N}(\mathbf{z}_i; \boldsymbol{\mu}_{cm}, \boldsymbol{\Sigma}_{cm})$ where $\mathbf{z}_i$ denotes the deep feature extracted from the i-th input, and $c_i \in \{1, \ldots, C\}$ is the corresponding ground-truth class label. The discriminative supervision is explicitly constructed as the sum of the multi-class Lovász-Softmax loss [7] and the focal loss [36], both computed using ground-truth semantic labels. The scalars $\lambda_g$ and $\lambda_d$ control the balance between generative feature modeling and discriminative segmentation performance. Importantly, GMMSeg treats the GMM parameters $\mu$ and $\Sigma$ as deterministic point estimates, limiting its ability to estimate the uncertainty over model parameters.

### 3.3 Epistemic Uncertainty Estimation

Unlike conventional GMM-based classifiers using point estimates, we treat the GMM parameters $\boldsymbol{\mu}_{cm}$ and $\boldsymbol{\Sigma}_{cm}$ as random variables with their own distributions. This allows us to express uncertainty over the parameters themselves. Note that this uncertainty is defined at the level of the GMM classifier head and is not interpreted as a full Bayesian treatment of the segmentation backbone. Formally, we place a conjugate prior over the mean and covariance of each Gaussian. Assuming that a feature vector assigned to a component $(c, m)$ follows a Gaussian distribution (we omit the subscripts $cm$ for brevity), $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the conjugate prior on the component parameters is the Normal–Inverse–Wishart (NIW) distribution [19, 45]: $\boldsymbol{\mu}|\boldsymbol{\Sigma} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}/\kappa_0)$, $\boldsymbol{\Sigma} \sim \mathrm{InvWishart}(\boldsymbol{\Psi}_0, \nu_0)$, where $\boldsymbol{\mu}_0 \in \mathbb{R}^D$ and $\boldsymbol{\Psi}_0 \in \mathbb{R}^{D \times D}$ denote prior location and scale, while $\kappa_0 > 0$ and $\nu_0 > D - 1$ control the prior precision and degrees of freedom.

Given an observed dataset $\mathcal{D}$ of $N$ feature vectors, with empirical mean $\bar{\mathbf{z}}$ and empirical covariance $\mathbf{S}$, the resulting posterior distribution will also be a NIW distribution: $(\boldsymbol{\mu}, \mathbf{\Sigma}|\mathcal{D}) \sim \mathrm{NIW}(\hat{\boldsymbol{\mu}}, \hat{\kappa}, \hat{\nu}, \hat{\boldsymbol{\Psi}})$, where the posterior parameters are updated analytically as:

$$\begin{aligned} \hat{\kappa} &= \kappa_0 + N, \quad \hat{\boldsymbol{\mu}} = \frac{\kappa_0 \boldsymbol{\mu}_0 + N\bar{\mathbf{z}}}{\kappa_0 + N}, \\ \hat{\nu} &= \nu_0 + N, \quad \hat{\boldsymbol{\Psi}} = \boldsymbol{\Psi}_0 + \mathbf{S} + \frac{\kappa_0 N}{\kappa_0 + N}(\bar{\mathbf{z}} - \boldsymbol{\mu}_0)(\bar{\mathbf{z}} - \boldsymbol{\mu}_0)^\top. \end{aligned} \tag{3}$$

This posterior defines a higher-order distribution over the GMM component parameters, capturing the epistemic uncertainty associated with the class-conditional likelihood $p(\mathbf{z}|c)$. Intuitively, the updated hyperparameters $(\hat{\kappa}, \hat{\nu})$ can be interpreted as evidence accumulated from $N$ observed samples, reflecting how confidently the model estimates each component's mean and covariance.

In the special case where feature dimensions are modeled independently, the multivariate Gaussian reduces to a product of univariate Gaussians. The corresponding conjugate prior becomes the Normal–Inverse–Gamma (NIG) distribution [19, 45], where for each scalar feature $z \in \mathbb{R}$, $(\mu, \sigma^2) \sim \mathrm{NIG}(\mu_0, \kappa_0, \alpha_0, \beta_0)$, and the posterior hyperparameters are updated as:

$$\begin{aligned} \hat{\kappa} &= \kappa_0 + N, \quad \hat{\mu} = \frac{\kappa_0 \mu_0 + N\bar{z}}{\kappa_0 + N}, \\ \hat{\alpha} &= \alpha_0 + \frac{N}{2}, \quad \hat{\beta} = \beta_0 + \frac{1}{2}\sum_{i=1}^{N}(z_i - \bar{z})^2 + \frac{\kappa_0 N}{2(\kappa_0 + N)}(\bar{z} - \mu_0)^2. \end{aligned} \tag{4}$$

In a standard full-covariance setting governed by NIW priors, the predictive distribution takes the form of a multivariate Student-t. In this work, under the assumption of feature independence, employing NIG priors, the multivariate posterior predictive density factorizes into a product of $D$ independent univariate Student-t distributions. For a specific component $m$ of class $c$, this is given by:

$$\begin{aligned} p(\mathbf{z}|c,m,\mathcal{D}) &= \iint p(\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\Sigma})p(\boldsymbol{\mu},\boldsymbol{\Sigma}|c,m,\mathcal{D})d\boldsymbol{\mu}d\boldsymbol{\Sigma} \\ &= \prod_{d=1}^{D} \mathcal{T}(z^{(d)}|\hat{\mu}_{cm}^{(d)}, \frac{(\hat{\kappa}_{cm}^{(d)}+1)\hat{\beta}_{cm}^{(d)}}{\hat{\kappa}_{cm}^{(d)}\hat{\alpha}_{cm}^{(d)}}, 2\hat{\alpha}_{cm}^{(d)}) \end{aligned} \tag{5}$$

Mixing all Gaussian components of the c-th class GMM, the class-conditional likelihood $p(\mathbf{z}|c;\boldsymbol{\phi}_c)$ reads: $p(\mathbf{z} \mid c) = \sum_{m=1}^{M} \pi_{cm}\, p(\mathbf{z} \mid c, m, \mathcal{D})$ and applying Bayes' rule (Eq. (1)) gives the corresponding posterior $p(c|\mathbf{z})$, which reflects uncertainty in the underlying generative model. Finally, predictive uncertainty is quantified as the Shannon entropy of the predictive distribution:

$$\mathcal{H}(c|\mathbf{z}) = -\sum_{c=1}^{C} p(c|\mathbf{z}) \log p(c|\mathbf{z}). \tag{6}$$

## 4 Experimental Setup

**Datasets and network architectures** We evaluate our proposed method in LiDAR semantic segmentation on the official validation split of SemanticKITTI [3] and nuScenes [10], which provide dense point-wise semantic annotations for LiDAR scans collected in diverse real-world driving scenarios. Each 3D scan is projected into a range-view image of size $64(H)\times1024(W)\times5$ for SemanticKITTI and $32(H) \times 1024(W) \times 5$ for nuScenes with five input features, which are annotated in 20 and 16 classes, respectively.

We employ three backbones that have shown the state-of-the-art LiDAR semantic segmentation performance: RangeFormer [27], a hierarchical transformer-based architecture, RangeViT [2], a generic Vision Transformer (ViT) architecture and SalsaNext [13], a CNN-based model with a U-Net-style architecture. All models extract features of dimension $D = 64$ and are trained using the multi-class loss function from Eq. (2), where $\lambda_g = 0.1$ and $\lambda_d = 0.9$. We use the AdamW [38] optimizer with an initial learning rate of 0.01 and utilize OneCycle learning rate scheduler [53] to adaptively adjust the learning rate during training. The models are trained with a batch size of 4 for 50 epochs on SemanticKITTI and 80 epochs on nuScenes. The number of Gaussian components per class is treated as a hyperparameter and is set to $M = 3$ in all experiments to optimize both segmentation and calibration, following the ablation study in Sec. 6.1.

To address the generalization of the proposed uncertainty head beyond range-view architectures, we additionally evaluate GMM-NIG on two more recent 3D segmentation backbones: PTv3 [57] and SphereFormer [30]. PTv3 is a point-based transformer architecture, while SphereFormer is a sparse-voxel/spherical transformer architecture. For these experiments, we attach the GMM-NIG head to the penultimate feature representation of each backbone.

### 4.1 Comparative Methods

We evaluate our method against established uncertainty estimation baselines. For deterministic approaches, we consider Maximum Softmax Probability (MSP) entropy [23], both uncalibrated and with post-hoc temperature scaling [22], Dirichlet scaling [29], and depth-aware scaling [28]. For the sampling-based family, we evaluate MC Dropout (10 stochastic forward passes) [17] and DE (5 independently trained models) [31]. Among sampling-free distribution-aware approaches, we include EDL [49]. Furthermore, to validate our contributions, we compare the GMMSeg [34] in its original point-estimation form with our Bayesian GMM-NIG formulation.

### 4.2 Evaluation Metrics

We assess the calibration of the predictive distribution using the Adaptive Calibration Error (ACE), which is the discrepancy between predicted confidence and empirical accuracy over adaptively defined bins as: $\mathrm{ACE} = \frac{1}{K}\sum_{k=1}^{K}|conf_k - acc_k|$, where $K$ is the number of non-empty bins, $\mathrm{conf}_k$ and $\mathrm{acc}_k$ are the average confidence and accuracy within bin $k$, respectively. Reliability diagrams are further used to analyze over- and underconfidence behavior. Segmentation performance is measured by mean Intersection over Union (mIoU). Failure detection is formulated as a binary task (correct vs. incorrect predictions) and evaluated using Area Under the Receiver Operating Characteristic Curve (AUROC) and Area Under the Precision-Recall Curve (AUPR).

## 5 Experimental Results

We conduct a comprehensive empirical evaluation to support the three main claims presented in Sec. 1, corresponding respectively to Sec. 5.1, Sec. 5.2, and Sec. 5.4. To support our fourth claim, we simultaneously consider both CNN-based and Transformer-based backbones in all experiments, demonstrating the backbone-agnostic nature of our approach. Additionally, qualitative visualizations (Sec. 5.5) further confirm that predicted uncertainty aligns with semantically ambiguous regions. We further investigate factorized and low-rank GMM covariance structures for modeling class-conditional distributions (see Sec. 5.6). Based on this analysis, the experiments adopt the factorized formulation due to its favorable trade-off between performance and computational efficiency.

### 5.1 Segmentation Accuracy

To support our first claim regarding the improved segmentation accuracy, Tab. 1 presents results on SemanticKITTI across three backbones. Overall, GMM-NIG consistently achieves the highest mIoU on all architectures, with DE as the second-best method. Specifically, for RangeFormer, mIoU improves from 63.60% (MSP) to 65.71% (+2.11% gain), while DE provides a +1.31% gain at roughly 5 times higher inference cost and the additional overhead of training multiple models. Similar improvements are observed for RangeViT and SalsaNext, confirming that our proposed posterior predictive formulation improves segmentation accuracy over both deterministic and sampling-based approaches.

On nuScenes (Tab. 2), the same pattern holds: although DE improves over the deterministic baseline, GMM-NIG attains higher mIoU than DE while maintaining inference time comparable to the standard single deterministic approach.

**Table 1:** Quantitative comparison on the validation set of SemanticKITTI. Best results are highlighted in bold, second-best in blue, and non-real-time inference in red.

| Backbone | Uncertainty Estimation Method | Segmentation | Calibration | Failure Detection | | Runtime (s) ↓ |
|---|---|---|---|---|---|---|
| | | mIoU (%) ↑ | ACE (%) ↓ | AUROC (%) ↑ | AUPR (%) ↑ | |
| RangeFormer [27] [ICCV 2023] | MSP Entropy (uncalibrated) | 63.60 | 4.97 | 81.93 | 37.33 | $0.21 \pm 0.01$ |
| | Temperature Scaling | 63.60 | 4.28 | 82.24 | 37.91 | $0.22 \pm 0.02$ |
| | Dirichlet Scaling | 63.60 | 4.36 | 82.01 | 37.31 | $0.21 \pm 0.01$ |
| | Depth-aware Scaling | 63.60 | 4.11 | 82.36 | 38.01 | $0.23 \pm 0.01$ |
| | EDL | 63.51 | 4.01 | 81.80 | 37.77 | $0.25 \pm 0.01$ |
| | Deep Ensembles(DE) | 64.91 | 2.81 | 87.61 | 48.07 | $1.31 \pm 0.03$ |
| | MC Dropout | 64.18 | 3.21 | 84.90 | 46.61 | $2.18 \pm 0.03$ |
| | GMMSeg | 63.90 | 3.88 | 86.11 | 47.01 | $0.22 \pm 0.03$ |
| | GMM-NIG (Ours) | **65.71** | **2.01** | **89.01** | **53.33** | $0.23 \pm 0.03$ |
| RangeViT [2] [CVPR 2023] | MSP Entropy (uncalibrated) | 57.58 | 5.51 | 81.07 | 35.01 | $0.38 \pm 0.02$ |
| | Temperature Scaling | 57.58 | 4.77 | 82.03 | 35.22 | $0.38 \pm 0.01$ |
| | Dirichlet Scaling | 57.58 | 6.91 | 80.36 | 34.77 | $0.38 \pm 0.02$ |
| | Depth-aware Scaling | 57.58 | 4.61 | 82.05 | 35.31 | $0.39 \pm 0.01$ |
| | EDL | 57.33 | 5.83 | 81.02 | 36.03 | $0.41 \pm 0.01$ |
| | Deep Ensembles(DE) | 61.88 | 4.30 | 87.90 | 41.03 | $1.96 \pm 0.02$ |
| | MC Dropout | 56.58 | 3.88 | 87.21 | 40.31 | $3.61 \pm 0.03$ |
| | GMMSeg | 61.01 | 6.02 | 81.34 | 41.67 | $0.41 \pm 0.02$ |
| | GMM-NIG (Ours) | **62.88** | **2.68** | **88.56** | **45.36** | $0.40 \pm 0.03$ |
| SalsaNext [13] [ISVC 2023] | MSP Entropy (uncalibrated) | 55.21 | 6.03 | 81.60 | 38.70 | $0.23 \pm 0.02$ |
| | Temperature Scaling | 55.21 | 5.28 | 82.08 | 39.03 | $0.23 \pm 0.01$ |
| | Dirichlet Scaling | 55.21 | 5.71 | 81.74 | 38.71 | $0.22 \pm 0.02$ |
| | Depth-aware Scaling | 55.21 | 4.98 | 82.33 | 39.41 | $0.24 \pm 0.01$ |
| | EDL | 55.63 | 5.32 | 79.91 | 35.03 | $0.25 \pm 0.01$ |
| | Deep Ensembles(DE) | 57.83 | 3.33 | 81.77 | 41.78 | $1.45 \pm 0.02$ |
| | MC Dropout | 55.10 | 4.43 | 84.75 | 41.90 | $2.21 \pm 0.02$ |
| | GMMSeg | 57.01 | 6.67 | 84.11 | 41.00 | $0.23 \pm 0.02$ |
| | GMM-NIG (Ours) | **57.96** | **2.10** | **86.32** | **53.68** | $0.24 \pm 0.02$ |

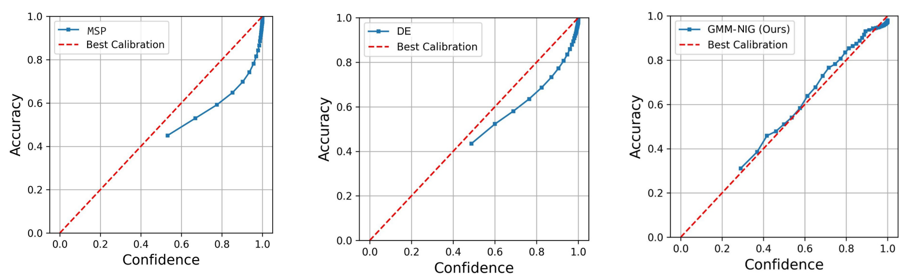
### 5.2 Confidence Calibration and Failure Detection

To address our second claim, we evaluate two aspects of the predictive distribution: (i) confidence calibration, where confidence is defined as the maximum posterior probability, and (ii) failure detection, utilizing the entropy of the predictive distribution as a proxy for uncertainty. The former assesses how well predicted probabilities align with the observed accuracy, while the latter evaluates how entropy correlates with prediction errors.

Overall, Tab. 1 shows that GMM-NIG consistently achieves the lowest calibration error across all backbones (2.01%, 2.68%, and 2.10%). Among the baselines, MSP yields the poorest calibration, while adding post-hoc methods helps

**Table 2:** Quantitative comparison on the validation set of the nuScenes dataset.

| Backbone | Uncertainty Estimation Method | Segmentation | Calibration | Failure Detection | | Time (s) ↓ |
|---|---|---|---|---|---|---|
| | | mIoU (%) ↑ | ACE (%) ↓ | AUROC (%) ↑ | AUPR (%) ↑ | |
| RangeFormer | MSP Entropy (uncalibrated) | 79.81 | 3.57 | 86.00 | 44.08 | $0.16 \pm 0.02$ |
| | Deep Ensembles(DE) | 80.11 | 2.01 | 88.90 | 51.50 | $0.98 \pm 0.03$ |
| | GMM-NIG (Ours) | **80.18** | **1.97** | **91.08** | **55.21** | $0.18 \pm 0.01$ |
| RangeViT | MSP Entropy (uncalibrated) | 73.81 | 3.71 | 84.21 | 40.33 | $0.25 \pm 0.1$ |
| | Deep Ensembles(DE) | 74.21 | 2.90 | 87.21 | 44.03 | $1.31 \pm 0.2$ |
| | GMM-NIG (Ours) | **75.88** | **1.26** | **87.40** | **47.01** | $0.31 \pm 0.1$ |
| SalsaNext | MSP Entropy (uncalibrated) | 69.99 | 6.03 | 83.00 | 39.56 | $0.15 \pm 0.1$ |
| | Deep Ensembles(DE) | 70.70 | 3.33 | 85.38 | 43.99 | $0.86 \pm 0.1$ |
| | GMM-NIG (Ours) | **71.23** | **2.10** | **86.10** | **54.21** | $0.18 \pm 0.2$ |




**(a)** Softmax Probability (MSP) **(b)** Deep Ensembles (DE) **(c)** GMM-NIG (ours)

**Fig. 2:** Reliability diagrams comparison (RangeViT backbone, SemanticKITTI data).

to reduce ACEs. Although DE and MC Dropout improve calibration by capturing epistemic uncertainty, they still result in larger ACEs compared to our approach. Notably, GMM-NIG outperforms its non-Bayesian counterpart (GMM-Seg), demonstrating the merit of modeling second-order distributions.

Fig. 2 compares reliability diagrams for the RangeViT backbone, contrasting our GMM-NIG model with MSP and the gold-standard DE. As illustrated in Fig. 2b, DE reduces the calibration gap compared to MSP, yet still exhibits a significant deviation from the ideal calibration line, indicating a higher calibration error and a clear tendency toward overconfidence. In contrast, our method, shown in Fig. 2c, demonstrates markedly improved calibration. The predicted confidence aligns more closely with the observed accuracy, yielding a lower ACE. Notably, our model tends to be slightly underconfident, reflecting a more conservative and reliable uncertainty estimation.

To evaluate the predictive uncertainty for failure detection, Tab. 1 summarizes the results on SemanticKITTI. Our method consistently outperforms the comparative methods across all backbones. On RangeFormer, AUROC improves from 87.61% (DE) to 89.01%, and AUPR increases from 48.07% to 53.33%. Comparable improvements are observed for RangeViT and in SalsaNext, only MC Dropout slightly outperforms DE. This highlights that our approach distinguishes erroneous predictions by assigning them higher uncertainty. The same behavior is consistently observed on nuScenes (Tab. 2).

**Table 3:** Generalization across LiDAR representation on SemanticKITTI.

| Backbone | Uncertainty Estimation Method | Segmentation | Calibration | Failure Detection | |
|---|---|---|---|---|---|
| | | mIoU (%) ↑ | ACE (%) ↓ | AUROC (%) ↑ | AUPR (%) ↑ |
| PTv3 [57] [CVPR'24] | MSP Entropy (uncalibrated) | 71.20 | 6.81 | 81.65 | 40.22 |
| | GMM-NIG (Ours) | **71.86** | **3.09** | **88.20** | **52.05** |
| SphereFormer [30] [CVPR'23] | MSP Entropy (uncalibrated) | 66.01 | 4.67 | 80.00 | 38.05 |
| | GMM-NIG (Ours) | **66.90** | **2.08** | **87.08** | **53.26** |

### 5.3 Generalization to 3D Backbones

To further address whether the proposed uncertainty head generalizes beyond the range-view backbones used in the main experiments, we evaluate GMM-NIG on PTv3 [57] which directly processes point-based representations and SphereFormer [30] which operates with sparse voxel features. As shown in Tab. 3, replacing the deterministic confidence estimate with the proposed GMM-NIG posterior predictive consistently reduces ACE by roughly 5% and improves failure prediction for both architectures, indicating that the benefit is not tied to a specific range-view representation or backbone design.

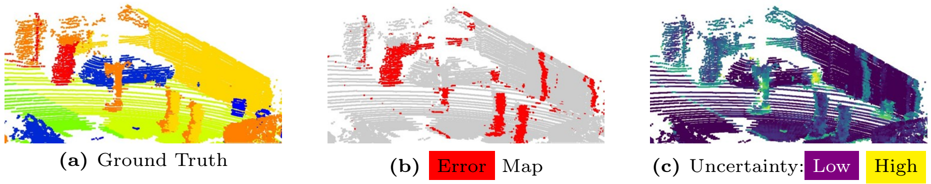
### 5.4 Performance vs. Efficiency

To validate our third claim, we compare the average inference time for a single LiDAR scan as shown in the last column of Tab. 1. Across all backbones, the computational cost results clearly reflect the overhead imposed by sampling-based methods. DE and MC Dropout are the least efficient approaches, with inference times increasing by an order of magnitude due to multiple forward passes (e.g., 1.96s and 3.61s for RangeViT compared to other methods around 0.23s). In contrast, our method requires only a single forward pass and achieves runtime comparable to deterministic models ($0.2-0.4$s per scan), while simultaneously surpassing both deterministic and sampling-based methods across all performance metrics. This combination of efficiency and superior uncertainty estimation makes it uniquely suited for real-time safety-critical applications.

While post-hoc methods and EDL offer superior inference efficiency, they provide limited improvement in confidence calibration and failure detection. Post-hoc techniques improve calibration without significantly enhancing failure detection, and EDL—despite its Dirichlet formulation—does not consistently surpass deterministic baselines in either metric.

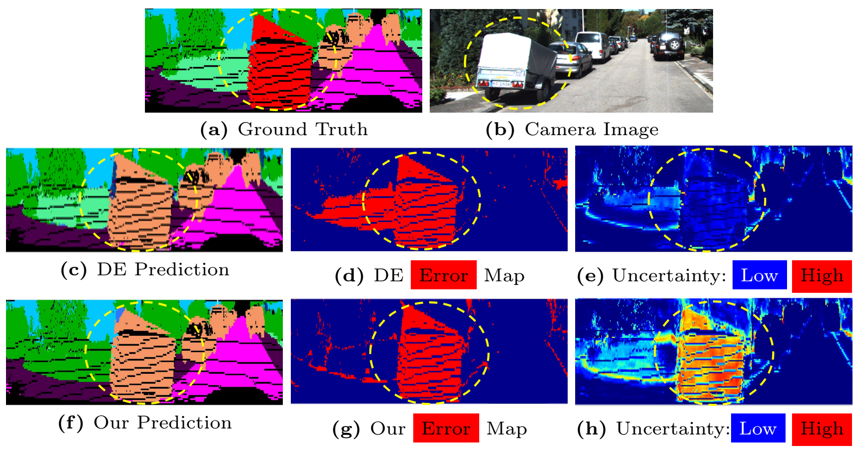
### 5.5 Qualitative Results

Fig. 3 illustrates the correlation between GMM–NIG's predictive uncertainty and actual errors. The uncertainty map (Fig. 3c) effectively localizes prediction failures, with high uncertainty mirroring the error map (Fig. 3b) at class boundaries and misclassified regions like buildings and trunks. We further compare GMM-NIG with DE on a representative range-view image in Fig. 4, highlighting



(a) Ground Truth (b) Error Map (c) Uncertainty: Low High

**Fig. 3:** GMM-NIG uncertainty vs. prediction errors for a representative LiDAR scan.



(a) Ground Truth (b) Camera Image

(c) DE Prediction (d) DE Error Map (e) Uncertainty: Low High

(f) Our Prediction (g) Our Error Map (h) Uncertainty: Low High

**Fig. 4:** Predictive uncertainty and actual errors (dashed yellow) for *GMM-NIG* vs. *Deep Ensembles.* car, other vehicle, road, sidewalk, building, vegetation, terrain

an example when DE fails to reflect uncertainty for the highlighted trailer misclassified as a car, whereas GMM-NIG assigns appropriately high uncertainty.

Qualitative analysis (including additional results in the supplementary material) reveals two consistent improvements. First, in ambiguous regions (e.g., parking–sidewalk or terrain–vegetation), DE and MC Dropout often produce overconfident misclassifications, whereas GMM-NIG better distinguishes these classes, supporting our first claim on accuracy. Second, in cases of inherent ambiguity where GMM-NIG also misclassifies (Fig. 4), it assigns significantly higher uncertainty to such predictions. In contrast, baselines often produce confident errors, validating our second claim regarding improved reliability.

### 5.6 Factorized vs. Low-rank Covariance

To evaluate the trade-off between modeling capacity and computational efficiency, we compare two covariance parameterizations for our GMM-based classifier: (i) a factorized formulation, modeling each feature dimension independently, and (ii) low-rank covariance approximation, where each class–component covariance matrix is represented in low-rank form following the Mixture of Factor

Analyzers model [20]: $\Sigma_{cm} = P_{cm}P_{cm}^{\top} + D_{cm}$, where $P_{cm} \in \mathbb{R}^{D\times R}$ is the factor loading matrix with rank $R$ (with $R$ being a hyperparameter and $R \ll D$), and $D_{cm} = \mathrm{diag}(\boldsymbol{\epsilon}_{cm}) \in \mathbb{R}^{D\times D}$ is a diagonal matrix representing the independent stochastic error of each feature dimension, which models the residual variability not explained by the low-rank factors.

**Table 4:** Quantitative comparison between factorized and low-rank Gaussian models.

| **Model** | **mIoU(%)**$\uparrow$ | **ACE(%)**$\downarrow$ | **FLOPs(G)**$\downarrow$ |
|---|---|---|---|
| GMM-NIG (factorized) | 57.96 | **2.10** | **63.16** |
| GMM-NIW (low-rank, R=8) | **58.01** | 2.87 | 66.18 |
| GMM-NIW (low-rank, R=12) | 57.33 | 2.81 | 67.94 |

For computational efficiency during the Sinkhorn-regularized E-step, we compute log-densities using the Woodbury matrix identity [55] and the matrix determinant lemma. In the M-step, the update is performed by projecting the weighted empirical covariance onto the low-rank structure via truncated eigendecomposition. Given hardware constraints (32 GB NVIDIA RTX 3060), we evaluated low-rank covariance models with ranks $R = 8$ and $R = 12$, and compared them against the factorized Gaussians using the SalsaNext backbone. Tab. 4 reports the segmentation accuracy (mIoU), calibration (ACE), and computational cost in FLOPs, measured for a single forward pass with batch size $B = 1$.

Overall, the factorized formulation achieves superior calibration, while the low-rank variants slightly improve accuracy at the cost of higher complexity. Specifically, the rank-8 model attains a marginal mIoU gain of 0.05% but exhibits a degradation in ACE (0.77%). Increasing the rank to $R = 12$ does not improve mIoU and instead worsens calibration. Both low-rank configurations incur additional computational overhead. In contrast, the factorized counterpart provides the best calibration and lowest FLOPs, representing the most efficient and well-calibrated configuration.

## 6 Ablation Study

### 6.1 Effect of the Number of Gaussian Components

We study the impact of varying the number of Gaussian components per class in the GMM-NIG model, represented in Tab. 5. Using a single Gaussian per class ($M = 1$) yields the lowest performance (54.3% mIoU, 3.04% ACE). Increasing $M$ to 3 substantially improves both segmentation accuracy and calibration (57.96% mIoU, 2.10% ACE), confirming the benefit of modeling class multimodality. However, further increasing $M$ to 5 or 10 yields only marginal or even negative gains, likely due to model overparameterization and reduced stability. Consequently, we set $M = 3$ as the default configuration, which offers a favorable balance between performance and computational efficiency. It is also worth noting that this hyperparameter affects both training and inference time. Additional ablation studies are provided in the supplementary material.

**Table 5:** Ablation study on the number of GMM components per class for SalsaNext.

| **Components** | **mIoU (%)** ↑ | **ACE (%)** ↓ |
|---|---|---|
| 1 | 54.3 | 3.04 |
| 3 | **57.96** | 2.10 |
| 5 | 56.21 | **2.01** |
| 10 | 56.33 | 2.86 |

## 7 Conclusion

We presented GMM-NIG, a Bayesian generative classifier that estimates sampling-free predictive uncertainty with a single deterministic forward pass. Following the GMMSeg generative classifier to model features for each class, we place conjugate NIG priors over class-conditional GMM parameters. By analytically marginalizing these parameters, our method replaces point-estimated Gaussians with a t-distribution posterior predictive, yielding calibrated confidence and a reliable uncertainty measure without the runtime and training overhead of MC Dropout or deep ensembles.

Across SemanticKITTI and nuScenes, and for both CNN (SalsaNext) and Transformer (RangeViT/RangeFormer) backbones, GMM-NIG consistently improves accuracy, achieves state-of-the-art calibration and failure detection, and matches the runtime of deterministic inference while avoiding the training and latency costs of sampling-based baselines.

Overall, GMM-NIG bridges the gap between fast discriminative segmentation and Bayesian uncertainty estimation by providing an analytic posterior predictive layer that is easy to integrate, robust across architectures, and effective on large-scale LiDAR benchmarks. It is broadly applicable to other classification tasks, as it makes no assumptions specific to LiDAR input data.

## Acknowledgements

This project is supported by the German Research Foundation (DFG), as part of the Research Training Group i.c.sens, GRK 2159, ‘Integrity and Collaboration in Dynamic Sensor Networks’.

## References


1. Ancha, S., Osteen, P.R., Roy, N.: Deep evidential uncertainty estimation for semantic segmentation under out-of-distribution obstacles. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 6943–6951. IEEE (2024)
2. Ando, A., Gidaris, S., Bursuc, A., Puy, G., Boulch, A., Marlet, R.: Rangevit: Towards vision transformers for 3d semantic segmentation in autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5240–5250 (2023)
3. Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., Gall, J.: Semantickitti: A dataset for semantic scene understanding of lidar sequences. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9297–9307 (2019)
4. Beluch, W.H., Genewein, T., Nürnberger, A., Köhler, J.M.: The power of ensembles for active learning in image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9368–9377 (2018)
5. Bengs, V., Hullermeier, E., Waegeman, W.: Pitfalls of epistemic uncertainty quantification through loss minimisation. Advances in Neural Information Processing Systems **35**, 29205–29216 (2022)
6. Bengs, V., Hüllermeier, E., Waegeman, W.: On second-order scoring rules for epistemic uncertainty quantification. In: International Conference on Machine Learning. pp. 2078–2091. PMLR (2023)
7. Berman, M., Triki, A.R., Blaschko, M.B.: The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4413–4421 (2018)
8. Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., West, M.: Generative or discriminative? getting the best of both worlds. Bayesian statistics **8**(3), 3–24 (2007)
9. Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D.: Weight uncertainty in neural network. In: International conference on machine learning. pp. 1613–1622. PMLR (2015)
10. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020)
11. Cheng, H.X., Han, X.F., Xiao, G.Q.: Cenet: Toward concise and efficient lidar semantic segmentation for autonomous driving. In: 2022 IEEE international conference on multimedia and expo (ICME). pp. 01–06. IEEE (2022)
12. Corbiere, C., Thome, N., Saporta, A., Vu, T.H., Cord, M., Perez, P.: Confidence estimation via auxiliary models. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(10), 6043–6055 (2021)
13. Cortinhal, T., Tzelepis, G., Erdal Aksoy, E.: Salsanext: Fast, uncertainty-aware semantic segmentation of lidar point clouds. In: Advances in Visual Computing: 15th International Symposium, ISVC 2020, San Diego, CA, USA, October 5–7, 2020, Proceedings, Part II 15. pp. 207–222. Springer (2020)
14. Dreissig, M., Piewak, F., Boedecker, J.: On the calibration of uncertainty estimation in lidar-based semantic segmentation. In: 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC). pp. 4798–4805. IEEE (2023)

15. Fan, L., Xiong, X., Wang, F., Wang, N., Zhang, Z.: Rangedet: In defense of range view for lidar-based 3d object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2918–2927 (2021)
16. Feng, D., Wei, X., Rosenbaum, L., Maki, A., Dietmayer, K.: Deep active learning for efficient training of a lidar 3d object detector. In: 2019 IEEE Intelligent Vehicles Symposium (IV). pp. 667–674. IEEE (2019)
17. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059. PMLR (2016)
18. Gawlikowski, J., Tassi, C.R.N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., et al.: A survey of uncertainty in deep neural networks. Artificial Intelligence Review **56**(Suppl 1), 1513–1589 (2023)
19. Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B.: Bayesian data analysis. Chapman and Hall/CRC (1995)
20. Ghahramani, Z., Hinton, G.E., et al.: The em algorithm for mixtures of factor analyzers. Tech. rep., Technical Report CRG-TR-96-1, University of Toronto (1996)
21. Graves, A.: Practical variational inference for neural networks. Advances in neural information processing systems **24** (2011)
22. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International conference on machine learning. pp. 1321–1330. PMLR (2017)
23. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv preprint arXiv:1610.02136 (2016)
24. Herron, E.J., Young, S.R., Potok, T.E.: Ensembles of networks produced from neural architecture search. In: International Conference on High Performance Computing. pp. 223–234. Springer (2020)
25. Juergens, M., Meinert, N., Bengs, V., Hüllermeier, E., Waegeman, W.: Is epistemic uncertainty faithfully represented by evidential deep learning methods? arXiv preprint arXiv:2402.09056 (2024)
26. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? Advances in neural information processing systems **30** (2017)
27. Kong, L., Liu, Y., Chen, R., Ma, Y., Zhu, X., Li, Y., Hou, Y., Qiao, Y., Liu, Z.: Rethinking range view representation for lidar segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 228–240 (2023)
28. Kong, L., Xu, X., Cen, J., Zhang, W., Pan, L., Chen, K., Liu, Z.: Calib3d: Calibrating model preferences for reliable 3d scene understanding. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1965–1978. IEEE (2025)
29. Kull, M., Perello Nieto, M., Kängsepp, M., Silva Filho, T., Song, H., Flach, P.: Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. Advances in neural information processing systems **32** (2019)
30. Lai, X., Chen, Y., Lu, F., Liu, J., Jia, J.: Spherical transformer for lidar-based 3d recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17545–17555 (2023)
31. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in neural information processing systems **30** (2017)
32. Lee, S., Purushwalkam, S., Cogswell, M., Crandall, D., Batra, D.: Why m heads are better than one: Training a diverse ensemble of deep networks. arXiv preprint arXiv:1511.06314 (2015)

33. Li, R., Li, S., Chen, X., Ma, T., Gall, J., Liang, J.: Tfnet: Exploiting temporal cues for fast and accurate lidar semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4547–4556 (2024)
34. Liang, C., Wang, W., Miao, J., Yang, Y.: Gmmseg: Gaussian mixture based generative semantic segmentation models. Advances in Neural Information Processing Systems **35**, 31360–31375 (2022)
35. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. arXiv preprint arXiv:1706.02690 (2017)
36. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
37. Liu, J., Lin, Z., Padhy, S., Tran, D., Bedrax Weiss, T., Lakshminarayanan, B.: Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. Advances in neural information processing systems **33**, 7498–7512 (2020)
38. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
39. Meinert, N., Gawlikowski, J., Lavin, A.: The unreasonable effectiveness of deep evidential regression. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 9134–9142 (2023)
40. Mena, G., Nejatbakhsh, A., Varol, E., Niles-Weed, J.: Sinkhorn em: an expectation-maximization algorithm based on entropic optimal transport. arXiv preprint arXiv:2006.16548 (2020)
41. Miandashti, H.S., Zou, Q., Brenner, C.: Uncertainty-aware scene understanding via efficient sampling-free confidence estimation. In: Conference on Robot Learning. pp. 4401–4417. PMLR (2025)
42. Milioto, A., Vizzo, I., Behley, J., Stachniss, C.: Rangenet++: Fast and accurate lidar semantic segmentation. In: 2019 IEEE/RSJ international conference on intelligent robots and systems (IROS). pp. 4213–4220. IEEE (2019)
43. Mohan, R., Hindel, J., Drews, F., Gläser, C., Cattaneo, D., Valada, A.: Open-set lidar panoptic segmentation guided by uncertainty-aware learning. arXiv preprint arXiv:2506.13265 (2025)
44. Mukhoti, J., Kirsch, A., Van Amersfoort, J., Torr, P.H., Gal, Y.: Deep deterministic uncertainty: A new simple baseline. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24384–24394 (2023)
45. Murphy, K.P.: Conjugate bayesian analysis of the gaussian distribution. def $\mathbf{1}(2\sigma 2)$, 16 (2007)
46. Ng, A., Jordan, M.: On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. Advances in neural information processing systems **14** (2001)
47. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., Snoek, J.: Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. Advances in neural information processing systems **32** (2019)
48. Ritter, H., Botev, A., Barber, D.: A scalable laplace approximation for neural networks. In: 6th international conference on learning representations, ICLR 2018-conference track proceedings. vol. 6. International Conference on Representation Learning (2018)
49. Sensoy, M., Kaplan, L., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. Advances in neural information processing systems **31** (2018)

50. Shen, M., Ryu, J.J., Ghosh, S., Bu, Y., Sattigeri, P., Das, S., Wornell, G.: Are uncertainty quantification capabilities of evidential deep learning a mirage? Advances in Neural Information Processing Systems **37**, 107830–107864 (2024)
51. Shojaei Miandashti, H., Zou, Q., Mehltretter, M.: Uncertainty estimation and out-of-distribution detection for lidar scene semantic segmentation. In: European Conference on Computer Vision. pp. 116–131. Springer (2024)
52. Sirohi, K., Marvi, S., Büscher, D., Burgard, W.: Uncertainty-aware lidar panoptic segmentation. arXiv preprint arXiv:2210.04472 (2022)
53. Smith, L.N., Topin, N.: Super-convergence: Very fast training of neural networks using large learning rates. In: Artificial intelligence and machine learning for multi-domain operations applications. vol. 11006, pp. 369–386. SPIE (2019)
54. Variani, E., McDermott, E., Heigold, G.: A gaussian mixture model layer jointly optimized with discriminative features within a deep neural network architecture. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4270–4274. IEEE (2015)
55. Woodbury, M.A.: Inverting modified matrices. Department of Statistics, Princeton University (1950)
56. Wu, B., Zhou, X., Zhao, S., Yue, X., Keutzer, K.: Squeezesegv2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud. In: 2019 international conference on robotics and automation (ICRA). pp. 4376–4382. IEEE (2019)
57. Wu, X., Jiang, L., Wang, P.S., Liu, Z., Liu, X., Qiao, Y., Ouyang, W., He, T., Zhao, H.: Point transformer v3: Simpler faster stronger. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4840–4851 (2024)
58. Xu, C., Wu, B., Wang, Z., Zhan, W., Vajda, P., Keutzer, K., Tomizuka, M.: Squeezesegv3: Spatially-adaptive convolution for efficient point-cloud segmentation. In: European Conference on Computer Vision. pp. 1–19. Springer (2020)
59. Xu, X., Kong, L., Shuai, H., Liu, Q.: Frnet: Frustum-range networks for scalable lidar segmentation. IEEE Transactions on Image Processing (2025)
60. Yu, L., Yang, B., Wang, T., Li, K., Chen, F.: Predictive uncertainty quantification for bird's eye view segmentation: A benchmark and novel loss function. arXiv preprint arXiv:2405.20986 (2024)
61. Zhao, Y., Bai, L., Huang, X.: Fidnet: Lidar point cloud semantic segmentation with fully interpolation decoding. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 4453–4458. IEEE (2021)