Evaluating Uncertainty Quantification for Bird's Eye View Semantic Segmentation

Bowen Yang[1], Linlin Yu[2], Tianhao Wang[2], Chagnbin Li[2], Feng Chen [2]
[1] Cypress Woods High School, bowenymail@gmail.com
[2] The University of Texas at Dallas, {linlin.yu, tianhao.wang, changbin.li, feng,chen}@utdallas.edu

ABSTRACT

The fusion of raw features from multiple sensors (such as stereo cameras, LiDAR, radars) equipped on an autonomous vehicle to create a Bird's Eye View (BEV) representation is an essential and powerful component in planning and controlling systems. Recently, there has been a significant surge of interest in utilizing deep learning models for BEV semantic segmentation. However, being able to anticipate errors in segmentation and improve the explainability of DNNs is crucial to autonomous driving. In this paper, we evaluate various uncertainty quantification methods for BEV semantic segmentation based on two benchmark datasets (CARLA and nuScenes) with two representative backbones (Lift-Splat-Shoot and cross-view transformer). We perform an extensive evaluation of several uncertainty quantification methods. Among these methods, the evidential and postnet methods consistently demonstrate better performance in uncertainty quantification compared to MC dropout, ensemble, and deterministic baseline methods. Additionally, the ensemble method consistently exhibits the best performance in segmentation. We also propose augmenting uncertainty-aware BEV semantic segmentation models with supervised camera-view semantic segmentation features. Through extensive experiments, we consistently observe improvements in both the quality of BEV segmentation and the quality of uncertainty quantification. These findings suggest that exploring different types of supervision holds promise as a direction for enhancing uncertainty-aware BEV segmentation models.

The code used for this paper can be found here: https://github.com/bluffish/uq_bevss. An appendix containing extra details regarding the dataset, model architecture, and plots to support reported results can be found at this link: https://shorturl.at/efACD.

1 INTRODUCTION

Deep neural networks (DNN) have had tremendous success in a large number of fields, including computer vision, natural language processing, and others. However, DNNs tend to make overconfident predictions on unseen, and unknown data and underconfident predictions on noisy data. The difficulty in verifying the correctness of DNNs has led to erroneous edge-case behaviors and other unexpected consequences of autonomous vehicles (AV).

There are two main types of uncertainty: *Epistemic uncertainty* is caused by a lack of knowledge, e.g. the input has regions very different from those observed in the training dataset and *aleatoric uncertainty* refers to uncertainty caused by randomness, such as noisy data and labels.

The first and classical family quantifies the predictive uncertainty of a DNN via multiple forward passes, such as deep ensemble [7] and dropout-based [4] methods. However, the requirement of excessive memory and computation burden makes them impossible for real-time applications. As proposed in recent years, the second family quantifies uncertainty using deterministic single forward-pass neural networks, including density-based and conjugate-prior-based methods. The density-based methods, such as posterior networks (postnet) [2], typically fit a distribution in the representation space and then use the associated PDF function to quantify different uncertainty types. The conjugate-prior-based methods, such as evidential neural networks (evidential) [14], train a deterministic neural network that directly predicts the conjugate prior distribution of the class probabilities for uncertainty quantification.

In this study, we investigate the problem of uncertainty quantification for BEV semantic segmentation (BEVSS). BEVSS has received increased attention in recent years and has been adopted in in modern AV systems, such as the Tesla Autopilot system. BEVSS aims to segment and categorize objects or regions in a top-down view of a scene based on the fusion of images from multiple cameras mounted on an AV. It involves predicting the class of each pixel in the BEV view, such as roads, lane marks, vehicles, etc. To the best of our knowledge, this is the first work evaluating various uncertainty quantification methods for BEVSS. Our main contributions are as follows:

- We present an extensive benchmark for evaluating uncertainty quantification for BEV semantic segmentation. This benchmark examines the performance of five representative uncertainty quantification methods (softmax entropy, deep ensemble, dropout-based, evidential, and postnet-based) on two benchmark datasets (CARLA and nuScenes) with different road maps and weather conditions. These uncertainty quantification methods are evaluated on two representative backbones (Lift-Splat-Shoot [12] and Cross-Viasdew Transformer [18]).
- Our empirical results demonstrate that evidential and postnet are consistently the most effective in quantifying aleatoric uncertainty for the task of misclassification detection. It is clear that there is room for large improvement due to the low AUPR scores for the task of misclassification detection. We also find that current methods are not effective for OOD detection, indicating pixel-wise OOD detection on BEV may be a difficult task.
- We demonstrate that supervising camera-view segmentation improves the quality of BEV segmentation and improves uncertainty quantification for all baselines and backbones by 1-4% on misclassification task. This result indicates a promising direction to explore different methods of supervision to improve the performance of existing methods for BEVSS.

2 BEV SEMANTIC SEGMENTATION

Existing BEVSS methods fall into two main categories: Lift splat shoot (LSS) [6, 10, 12] and transformer-based approaches [11, 18]. This paper uses two representative methods, LSS [12] and Cross-View Transformer (CVT) [18], as the backbones of our study.

LSS leverages raw pixel inputs from multiple surrounding cameras and "lifts" each image individually into a frustum of features. Initially, it predicts a categorical distribution over a predefined set of possible depths. Subsequently, the frustum of features is generated by multiplying the features with their predicted depth probability. By utilizing known camera calibration matrices for each camera, a point cloud of features in the ego coordinate space can be obtained. LSS then "splats" all the frustums into a rasterized bird's-eye-view grid using a PointPillar [8] model. These splatted features are then fed into a decoder module to predict BEVSS. The concept of transforming from camera pixels to 3D point clouds and subsequently to BEV pixels has inspired several subsequent models, such as BEVDet [6] and FIERY [5].

CVT takes a distinct approach by leveraging transformer architecture and cross-attention mechanism. CVT begins by extracting features from multiple surrounding camera images using a pre-trained EfficientNet-B4[17] model. These extracted features serve as the attention values in the subsequent cross-attention step. To create the attention keys, the features are concatenated with the camera-aware positional embedding. This positional embedding is constructed using known camera pose and intrinsic information, enabling the model to account for the specific characteristics of each camera. The positional encoding of the BEV space serves as the queries during the cross attention process.

3 UNCERTAINTY QUANTIFICATION ON BEV SEMANTIC SEGMENTATION

Suppose we are given n images from different RGB camera views of the ego vehicle. Let $\mathbf{X} := \{\mathbf{X}_k, \mathbf{E}_k, \mathbf{I}_k\}_{k=1}^{n}$ denote the input, where each camera view has a feature matrix $\mathbf{X}_k \in \mathbb{R}^{3\times H\times W}$, extrinsic matrix $\mathbf{E}_k \in \mathbb{R}^{3\times 4}$, and intrinsic matrix $\mathbf{I}_k \in \mathbb{R}^{3\times 3}$. BEV semantic segmentation aims to predict the pixel-level classes in the BEV coordinate frame $\mathbf{Y} \in \{0,1\}^{C\times X\times Y}$, where C is the number of classes and X and Y are width and height of the BEV frame, respectively. A BEV neural network has the general form: $\mathbf{P} = f(\mathbf{X};\boldsymbol{\theta})$, where $\mathbf{P} \in [0,1]^{C\times X\times Y}$ are the pixel-wise class probabilities and $\boldsymbol{\theta}$ refers to the nework weights. We use $\mathbf{p}_{i,j}$ to denote the class-probability vector of the BEV pixel indexed by (i,j).

We will investigate different uncertainty quantification methods to quantify the aleatoric uncertainty $u_{i,j}^{alea}$ and the epistemic uncertainty $u_{i,j}^{epis}$ of the BEV network for each BEV pixel (i,j). We will consider the following representative methods: softmax, deep ensemble-based, dropout-based, evidential neural networks, and postnet-based methods.

Softmax-based. Softmax entropy is one of the most commonly used metrics for uncertainty. It is the entropy ($\mathbb{H}(p(\mathbf{Y}_{i,j}|\mathbf{X};\boldsymbol{\theta}))$) of softmax distribution $p(\mathbf{Y}_{i,j}|\mathbf{X};\boldsymbol{\theta})$.

$$\mathbb{H}(p(\mathbf{Y}_{i,j}|\mathbf{X};\boldsymbol{\theta})) = -\sum_{c=1}^{C} P_{c,i,j}\log P_{c,i,j}. \tag{1}$$

This metric is known to capture aleatoric uncertainty, but can not capture epistemic uncertainty reliably.

Deep-ensembles-based [7]. Deep Ensembles-based method learns M different versions of network weights $\{\boldsymbol{\theta}^{(1)},\cdots,\boldsymbol{\theta}^{(M)}\}$ based on different random seeds and aggregates the predictions of these versions. The aleatoric uncertainty is measured by the softmax entropy of the mean of the predictions from different network weights. The epistemic uncertainty is measured the variance between the model predictions.

$$\begin{aligned} u_{i,j}^{alea} &= -\sum_{c=1}^{C}\left(\frac{1}{M}\sum_{m=1}^{M} P_{c,i,j}^{(m)}\right)\log\left(\frac{1}{M}\sum_{m=1}^{M} P_{c,i,j}^{(m)}\right) \\ u_{i,j}^{epis} &= \frac{1}{\sum_c \mathrm{var}(\{P_{c,i,j}^{(m)}\}_{m=1}^{M})}, \end{aligned} \tag{2}$$

where $\mathbf{P}^{(m)} \in [0,1]^{C\times X\times Y}$ refers to the predictions of BEV network based on the network weights $\boldsymbol{\theta}^{(m)}$.

Dropout-based [4]. It is a method that approximates Bayesian inference based on dropout at test time. It conducts multiple stochastic forward passes with active dropout layers at test time. Similar to Deep ensembles, the entropy of expected softmax probability and variance of multiple predictions are used as uncertainty scores.

Evidential [14] Evidential neural network (ENN) was originally designed for singe-output classification tasks (e.g., image classification) and predicts the parameters of a Dirichlet distribution instead of class probabilities (the parameters of a categorical distribution). Dirichlet is the conjugate prior to the categorical distribution (or second-order distribution of the class label) and can effectively quantify different types of uncertainty, such as epistemic and aleatoric uncertainty. In this study, we extend evidential to conduct uncertainty quantification for BEV semantic segmentation. We use the same BEV network architecture, except that we replace the softmax activation function with the ReLU activation function to predict the concentration parameters ($\boldsymbol{\alpha}_{i,j} \in \mathbb{R}^{+C}$) of a Dirichlet for each pixel (i,j) in the BEV frame.

$$Y_{i,j} \sim \mathrm{Cat}(\mathbf{p}_{i,j}),\ \mathbf{p}_{i,j} \sim \mathrm{Dir}(\boldsymbol{\alpha}_{i,j}). \tag{3}$$

The aleatoric and epistemic uncertainty of the pixel (i,j) can be estimated based on the Dirichlet parameters $\boldsymbol{\alpha}_{i,j}$:

$$u_{i,j}^{alea} = -\max_c \bar{p}_{c,i,j},\ u_{i,j}^{epis} = C/\sum_{c=1}^{C}\alpha_{c,i,j}, \tag{4}$$

where $\bar{\mathbf{p}}_{i,j} := \mathbb{E}[\mathbf{p}_{i,j}|\boldsymbol{\alpha}_{i,j}]$ is the expected class-probability vector calculated based on the Dirichlet distribution $\mathrm{Dir}(\boldsymbol{\alpha}_{i,j})$.

Postnet-based [2] This method is designed based on feature space density with the prior knowledge that instances with high epistemic uncertainty should be far away from the training data in the latent space. Postnet was originally designed for single-output classification tasks and later extended to uncertainty quantification for graph neural networks [16]. Postnet first applies an encoding network (such as VGG [15]) to map the raw features to a low-dimensional latent feature and then applies a normalizing flow layer to estimate the densities of data points on latent space. The estimated densities are used to calculate the parameters of a Dirichlet Distribution. To generalize Postnet to BEV semantic segmentation, we adopt the LSS or CVT as the encoding network and add convolutional layers to propagate the pixel-level predicted evidence among the spatially

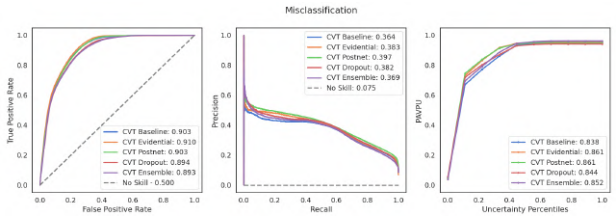

Figure 1: Misclassification detection result on CARLA for all Cross View Transformer-based models without segmentation: the left one shows the ROC curve, the middle one shows the PR curve and the right one shows the PAvPU plot based on different uncertainty thresholds. Numbers next to legend are area under curve values. Evidential and Postnet have the highest AUC values for most metrics

correlated pixels. As Postnet and evidential are both predicting Dirichlet distributions, we can use the same formulas in Eq. (4) to estimate uncertainties.

4 EXPERIMENTS

In this section, we evaluate several baselines for uncertainty quantification in BEVSS on two backbones (i.e. LSS and CVT) using extensive empirical evaluation on a dataset we generated using the CARLA simulator [3] and the nuScenes dataset[1].

4.1 Experimental Setup

Datasets We generate a synthetic dataset with CARLA simulator [3]. The data collection procedure is as follows: We collect 50k frames across 100 episodes and five ego vehicles in each episode. Six cameras are mounted to the top of each ego vehicle at different angles: -60, 0, 60, -120, 120, and 180. For each camera, we gather RGB images of the road scene with a resolution of 128×352 (no ego vehicle shown in the image) and the corresponding pixel-wise ground truth depth map, and ground truth semantic segmentation in the camera view. The ground-truth BEV segmentation is derived by a camera located 30m above the ground facing down with a resolution of 200×200. Our vehicles drive around in 15 weather conditions: Default, Clear Noon, Cloudy Noon, Wet Noon, Wet Cloudy Noon, Mid Rainy Noon, Hard Rain Noon, Soft Rain Noon, Clear Sunset, Cloudy Sunset, Wet Sunset, Wet Cloudy Sunset, Mid Rain Sunset, Hard Rain Sunset, and Soft Rain Sunset.

We also conduct a case study on the nuScenes dataset[1], which consists of image data from 1k scenes. Similar to the CARLA dataset, the ego vehicle has 6 cameras mounted at similar angles. Each scene lasts 20 seconds and consists of 40 frames of data. There are 40k total scenes of data. We use four classes (vehicle, road, lane, background) to evaluate the performance of BEV semantic segmentation.

Architecture For each uncertainty quantification baseline, we train two different backbones: Lift Splat Shoot, and Cross View Transformer. We also consider supervising camera-view segmentation using ground truth camera-view segmentation. We train UNet[13] for camera view semantic segmentation and then feed the prediction into the backbone as additional channels, that is $\mathbf{X}_k \in \mathbb{R}^{7\times H\times W}$. The first three channels are original RGB features. The next four are predicted logits for four target classes from UNet. We train the backbone and UNet together.

Training setup We train each model for around 15 epochs on the CARLA Dataset, and 30 epochs on nuScenes. Models with the Lift Splat Shoot backbone are trained using Adam as the optimizer, with a learning rate of 1e-3 and a weight decay of 1e-7. Models with the Cross View Transformer backbone are trained using AdamW as the optimizer and the one cycle learning rate scheduler, with a learning rate of 4e-3 and a weight decay of 1e-7. All models were trained with a batch size of 32, and trained using 2 or 4 GPUs.

Metrics for evaluation We use the *mean intersection over union (mIOU)* score to evaluate the quality of semantic segmentation for each class. The evaluation of uncertainty quantification performance is conducted using the *Patch Accuracy vs Patch Uncertainty* (PAvPU) metric proposed in [9], which indicates the probability of the model being confident in making accurate predictions or uncertain in the case of inaccurate predictions. We use the same hyperparameters as [9] including the patch size and accuracy threshold. In addition, misclassification detection and Out-of-Distribution (OOD) detection are typical evaluation tasks for assessing the effectiveness of uncertainty quantification in the image classification domain. These tasks involve binary classification, and performance are demonstrated using *Area Under Receiver Operating Characteristic* (AUROC), *Area Under Precision Recall* (AUPR) curves. We conduct the pixel-level misclassification detection task detecting whether a given prediction is incorrect using an uncertainty score, and the pixel-level Out-of-Distribution (OOD) detection task detecting whether an input pixel belongs to an OOD class given an estimate of uncertainty.

4.2 Results on CARLA

In Table 1, we report all quantitative results on 5 baseline uncertainty quantification methods based on two backbones (LSS and CVT), including mIOU for semantic segmentation, as well as PAvPU, AUROC, and PR for the misclassification detection task.

Semantic Segmentation performance Similar to observation in [12] (LSS), the 'Road' and the 'Background' segmentation prove to be much easier than 'Lane' and 'Vehicle' segmentation. Where the former can achieve over 90% mIOU, 'Lane' is below 50%. This is because lane detection necessitates a certain level of semantic understanding of the environment, while vehicle detection is hindered by the wide range of appearances that vehicles can exhibit. Additionally, both lanes and vehicles represent a smaller proportion of pixels within birds eye view, and inaccuracies in the mapping process will have a greater impact on the overall mIOU for those clasess. We also observed that CVT-based models achieve a higher mIOU than LSS-based models, from 8.77% percent to 15.2% percent on vehicle detection, as well as similar improvements on detection of other objects. As a more recent work, CVT uses a larger version of EfficientNet (EfficientNet-B4) as the image backbone compared to LSS (EfficientNet-B0), and the transformer architecture it uses already shown remarkable success in various fields. We observe that adding supervised segmentation improves mIOU for vehicle detection by 4.47% to 8.66% on LSS, and by 1.11% - 3.72% on CVT, as well as improvements in other classes.

Uncertainty Quantification Performance In the misclassification detection task, the aleatoric uncertainty score is employed for ranking, while the epistemic uncertainty score is used for the OOD detection task. For the misclassification task, the evidential and postnet have the best AUROC, AUPR, and AU-PAvPU. The deterministic baseline consistently performs the worst. Similar to

Table 1: Quantitative Results on CARLA dataset

Backbone	Baseline	Vehicle	Road	Lane	Background	Runtime (ms)	AUPR	AUROC	AU-PAvPU
		Mean IOU for Semantic Segmentation				Average	Misclassification		
Lift Splat shoot w/o Segmentation Supervision	Baseline	51.55%	89.96%	39.16%	97.68%	41.03ms	0.333	0.886	0.812
	Evidential	51.73%	90.17%	38.91%	97.68%	45.45ms	0.373	0.889	0.838
	Postnet	47.60%	89.80%	38.94%	97.68%	39.95ms	0.388	0.882	0.833
	Dropout	51.73%	90.12%	37.25%	97.68%	425.08ms	0.348	0.879	0.832
	Ensemble	60.96%	91.07%	42.74%	97.68%	146.84ms	0.357	0.896	0.832
Lift Splat shoot w/ Segmentation Supervision	Baseline	56.28%	90.37%	39.57%	97.74%	66.83ms	0.356	0.894	0.822
	Evidential	56.20%	90.41%	39.72%	97.79%	67.01mfs	0.382	0.897	0.848
	Postnet	56.26%	90.47%	39.11%	97.71%	72.34ms	0.392	0.899	0.848
	Dropout	56.30%	90.55%	39.36%	97.87%	455.77ms	0.359	0.881	0.828
	Ensemble	66.47%	91.42%	42.97%	97.40%	234.05ms	0.361	0.893	0.823
Cross View Transformer w/o Segmentation Supervision	Baseline	63.08%	91.09%	40.17%	98.11%	**25.34**ms	0.364	0.903	0.838
	Evidential	63.50%	91.20%	40.13%	98.12%	27.51ms	0.383	0.910	0.861
	Postnet	62.80%	90.98%	39.44%	98.11%	30.31ms	0.399	0.903	0.861
	Dropout	63.39%	91.35%	39.04%	98.09%	321.52ms	0.382	0.894	0.844
	Ensemble	69.73%	92.46%	45.57%	98.34%	104.90ms	0.367	0.893	0.852
Cross View Transformer w/ Segmentation Supervision	Baseline	65.07%	91.42%	42.04%	**98.39**%	52.43ms	0.389	0.936	0.884
	Evidential	66.29%	91.52%	41.86%	98.42%	51.99ms	**0.412**	0.926	0.892
	Postnet	64.35%	90.91%	40.04%	98.32%	66.33ms	0.410	**0.942**	0.892
	Dropout	64.50%	91.20%	40.21%	98.36%	594.86ms	0.387	0.933	0.890
	Ensemble	**73.45**%	**92.77**%	**47.63**%	98.32%	120.38ms	0.374	0.940	**0.893**

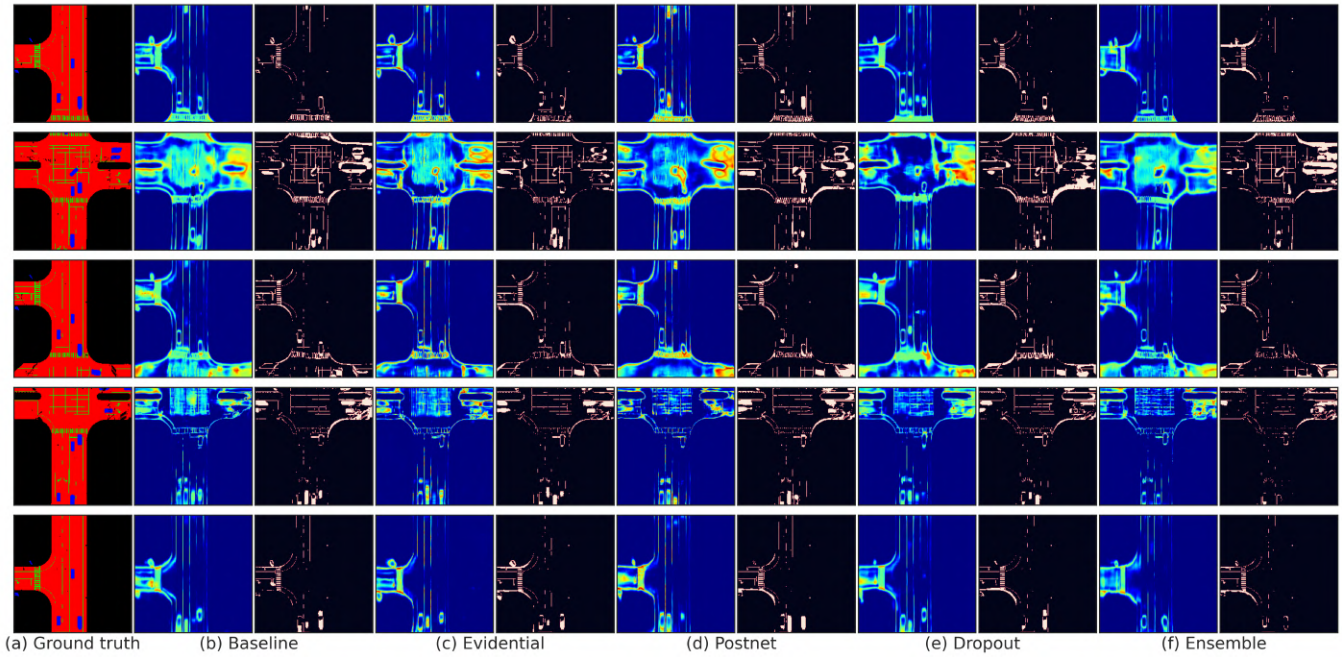

Figure 2: Visualization of aleatoric uncertainty baselines on samples from the CARLA validation set. The first column is the ground truth birds eye view semantic segmentation. Each pair of columns after that represent one baseline. The left column represents the birds eye view uncertainty map. The right column represents the misclassified pixels of the prediction from the baseline, with the lighter color meaning misclassified. The first pair of columns represent baseline, the second represent evidential, third - postnet, fourth - dropout, and fifth - ensemble.

semantic segmentation performance, CVT-based models perform better than LSS-based by 1-3% in AUROC, 3-5% AUPR, and 2-3% AU-PAvPU. Although all models yield less than satisfying performance comparing to well-studied task such as uncertainty quantification on image classification, we observe that supervising camera view segmentation consistently improves uncertainty quantification performance by 1-3% in AUROC, 2-4% in AUPR, and 2-4% in AU-PAvPU, indicating a promising direction to enhance these models. For the pixel-level OOD detection task, we observe a high AUROC but extremely low AUPR value, showing a bad performance with the current predicted epistemic uncertainty. We guess that this is because pixel-level OOD detection is quite a difficult task. First, pixel segmentation is more difficult than object detection. Also, BEV segmentation involves projecting information to a different view, increasing the difficulty.

We show a visualization of several samples from the CARLA dataset in Figure 2. Note that there is one ego car in each sample and models can not predict it in the BEV due to the lack of its own information in the camera view. We can see those object boundaries are easily misclassified and tend to have higher predicted aleatoric uncertainty, indicating effective aleatoric uncertainty estimation for all models. Among all uncertainty quantification methods, regions of high uncertainty predictions from the deterministic baseline are not quite consistent with the misclassification map, where multiple correctly classified areas have high aleatoric uncertainty, indicating bad calibration. Evidential and postnet models tend to have sharper differences between neighboring pixels than ensemble and dropout models, which indicates a more concentrated and precise prediction of aleatoric uncertainty.

Notably, models that utilize a single forward pass, such as deterministic baseline, evidential, and postnet, exhibit significantly less average runtime when compared to models that involve multiple forward passes, such as dropout and ensemble. The average runtime for dropout is the highest because a single forward pass constitutes 20 predictions of the model and a forward pass for an ensemble consists of forward passes from 5 different models.

4.3 Case study on nuScenes dataset

We show the performance for uncertainty quantification on the nuScenes dataset with misclassification detection in Figure 3. We perform this evaluation using the Cross View Transformer backbone, as it has better results on the CARLA dataset. We observe that the evidential and postnet baselines outperform the other baselines, which is consistent with observations from CARLA dataset.

5 CONCLUSION

This paper presents a comprehensive evaluation of multiple uncertainty quantification methods for bird's eye view semantic segmentation. The evaluation is conducted through an empirical study using two different BEV backbones (LSS and CVT) and two distinct datasets (CARLA and nuScenes). Our findings indicate that across all baselines and datasets, the evidential and postnet methods for uncertainty quantification consistently outperform the other methods. However, we observe that none of the baselines achieve satisfactory results, particularly for the task of OOD detection, highlighting the need for significant improvement in this area. To address this limitation, we demonstrate the effectiveness of incorporating supervised camera-view segmentation, which consistently enhances the overall performance. This approach can be further extended to explore other supervision techniques as potential avenues for improving the models.

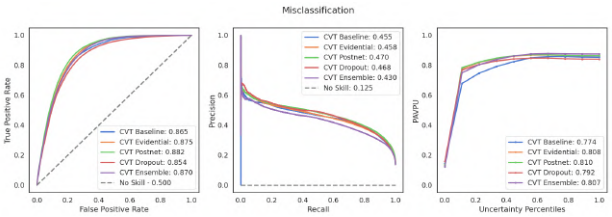

Figure 3: Misclassification detection result on the nuScenes dataset for all Cross View Transformer-based models without segmentation supervision: the left one shows the ROC curve, the middle one shows the PR curve and the right one shows the PAvPU plots based on different uncertainty thresholds. Numbers next to legend are area under curve values. Evidential and Postnet still have the highest AUC values for most metrics

REFERENCES

[1] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. 2020. nuScenes: A multimodal dataset for autonomous driving. arXiv:1903.11027 [cs.LG]

[2] Bertrand Charpentier, Daniel Zügner, and Stephan Günnemann. 2020. Posterior network: Uncertainty estimation without ood samples via density-based pseudo-counts. *Advances in Neural Information Processing Systems* 33 (2020), 1356–1367.

[3] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. 2017. CARLA: An Open Urban Driving Simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*. 1–16.

[4] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*. PMLR, 1050–1059.

[5] Anthony Hu, Zak Murez, Nikhil Mohan, Sofia Dudas, Jeffrey Hawke, Vijay Badrinarayanan, Roberto Cipolla, and Alex Kendall. 2021. FIERY: future instance prediction in bird's-eye view from surround monocular cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15273–15282.

[6] Junjie Huang, Guan Huang, Zheng Zhu, and Dalong Du. 2021. BEVDet: High-performance Multi-camera 3D Object Detection in Bird-Eye-View. *ArXiv* abs/2112.11790 (2021).

[7] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *NeurIPS* 30 (2017).

[8] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. 2018. PointPillars: Fast Encoders for Object Detection From Point Clouds. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2018), 12689–12697.

[9] Jishnu Mukhoti and Yarin Gal. 2018. Evaluating bayesian deep learning methods for semantic segmentation. *arXiv preprint arXiv:1811.12709* (2018).

[10] Mong H Ng, Kaahan Radia, Jianfei Chen, Dequan Wang, Ionel Gog, and Joseph E Gonzalez. 2020. BEV-Seg: Bird's Eye View Semantic Segmentation Using Geometry and Semantic Point Cloud. *arXiv preprint arXiv:2006.11436* (2020).

[11] Lang Peng, Zhirong Chen, Zhangjie Fu, Pengpeng Liang, and Erkang Cheng. 2022. BEVSegFormer: Bird's Eye View Semantic Segmentation From Arbitrary Camera Rigs. arXiv:2203.04050 [cs.CV]

[12] Jonah Philion and Sanja Fidler. 2020. Lift, Splat, Shoot: Encoding Images From Arbitrary Camera Rigs by Implicitly Unprojecting to 3D. In *ECCV*.

[13] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *ArXiv* abs/1505.04597 (2015).

[14] Murat Sensoy, Lance Kaplan, and Melih Kandemir. 2018. Evidential deep learning to quantify classification uncertainty. *NeurIPS* 31 (2018).

[15] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[16] Maximilian Stadler, Bertrand Charpentier, Simon Geisler, Daniel Zügner, and Stephan Günnemann. 2021. Graph posterior network: Bayesian predictive uncertainty for node classification. *Advances in Neural Information Processing Systems* 34 (2021), 18033–18048.

[17] Mingxing Tan and Quoc V. Le. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *ArXiv* abs/1905.11946 (2019).

[18] Brady Zhou and Philipp Krähenbühl. 2022. Cross-view transformers for real-time map-view semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13760–13769.