We thank all reviewers for their valuable comments. We provide additional experimental results here.

1 Additional Experiments

Ablation Study on LSS-based proposed framework on nuScenes: We add additional ablation study in Table 1.Specifically, we conducted an ablation study using the CVT backbone on the nuScenes dataset, which is detailed in the main paper.

In this material, we report results with LSS as the backbone. Initially, replacing our proposed UFCE loss with the standard UCE loss, which includes entropy regularization (ENT), leads to an increase in the Expected Calibration Error (ECE) score and higher misclassification AUROC and AUPR, indicating poorer calibration and reduced misclassification detection performance. Removing vacuity scaling shows no notable impact on pure segmentation, calibration, or misclassification detection; however, it results in over a 10% decline in OOD detection PR and 4% increase in AUROC. This suggests that while vacuity scaling reduces the ability to distinguish lower-ranked nodes, it largely aids in identifying true positives among the top-ranked nodes. This is particularly important in scenarios requiring manual verification of top-ranked misclassifications or anomalies, where a high rate of true positives is crucial, making AUPR a more appropriate metric than AUROC. Furthermore, eliminating both vacuity scaling and pseudo-OOD exposure (ER) significantly degrades OOD detection performance.

Thus, we can assert that our UFCE loss provides better aleatoric uncertainty prediction and calibration than the commonly used UCE loss across all configurations. Both vacuity scaling and pseudo-OOD exposure are essential for accurate epistemic uncertainty prediction.

Table 1: Ablation study on LSS-based proposed framework on nuScenes

Model	Pure Classification		Misclassification		OOD	
	Vehicle IoU	ECE	AUROC	AUPR	AUROC	AUPR
UFCE-VS-ER	**0.37**	**0.00145**	0.91	0.33	0.82	**0.31**
UCE-ENT-VS-ER	**0.37**	0.00502	0.874	0.315	0.868	0.206
UFCE-ER	0.339	0.0017	0.932	0.321	**0.929**	0.206
UFCE	0.366	0.00303	**0.945**	**0.336**	0.607	0.000554

Results on an additional Dataset: We conduct additional experiments on Lyft (Kesten et al., 2019), which is also used in LSS (Philion and Fidler, 2020) paper. Due to time constraints, these experiments were limited to evidential-based models.

Our findings on Lyft from Table 2 are consistent with those from our studies on CARLA and nuScenes. Overall, our proposed model employing UFCE loss consistently outperforms the UCE loss in terms of semantic segmentation, calibration, and misclassification detection. Moreover, incorporating vacuity scaling and pseudo-OOD exposure significantly enhances OOD detection performance.

Table 2: Ablation study on the LSS-based proposed framework on Lyft.

Model	Pure Classification		Misclassification		OOD	
	Vehicle IoU	ECE	AUROC	AUPR	AUROC	AUPR
UFCE-VS-ER	**0.474**	**0.00383**	**0.879**	**0.320**	0.766	**0.165**
UCE-ENT-VS-ER	0.426	0.00656	0.778	0.254	**0.842**	0.150
UFCE-ER	0.462	0.00580	0.872	0.310	0.758	0.114
UFCE	0.448	0.00926	0.831	0.289	0.518	0.00109
UCE-ENT	0.426	0.00516	0.763	0.241	0.552	0.00131

Model variance: We report the model variance for our proposed model in Table 3 and the evidential UCE baseline. We randomly initialize the model three times and the variance is within 3%.

Table 3: Variance for CVT on nuScenes.

Num. Models	Model	Pure Classification		Misclassification		OOD	
		Vehicle IoU	ECE	AUROC	AUPR	AUROC	AUPR
3	UFCE-VS-ER	0.344 ± 0.000013	0.000793 ± 0.00000016	0.944 ± 0.000009	0.328 ± 0.000021	0.924 ± 0.00021	0.287 ± 0.0029
	UCE-ENT-VS-ER	0.313 ± 0.0000053	0.00353 ± 0.0000000511	0.927 ± 0.0000023	0.318 ± 0.00001	0.875 ± 0.0031	0.254 ± 0.0000063

Qualitative Visualization: We qualitatively analyze the model performance with pixel-level prediction when viewed from a bird's eye perspective, and show the effectiveness of each component of our proposed model.

We first plot the semantic segmentation prediction in Figure 1. The full version of our model has the most precise segmentation results.

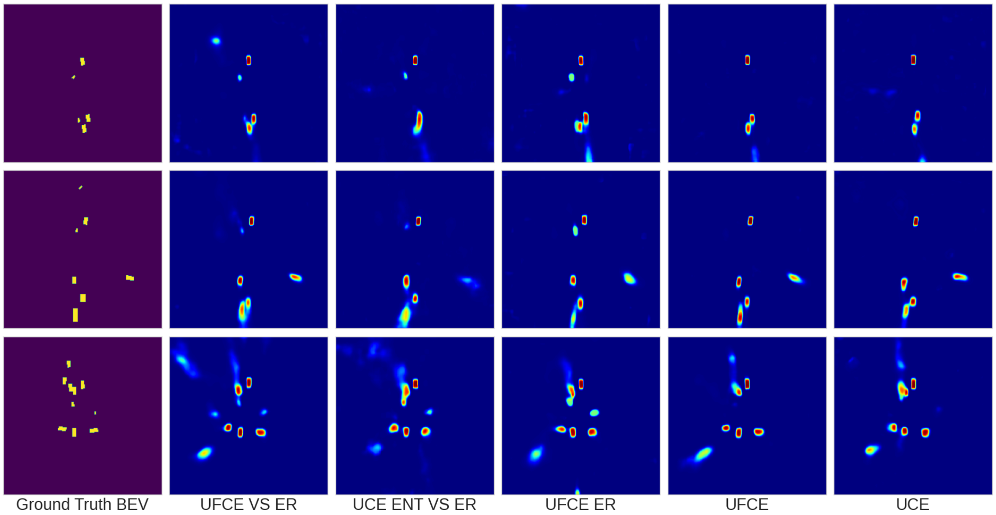

Figure 1: ViZ3

We present the predicted aleatoric uncertainty in Figure 2. We anticipate that correctly classified pixels will exhibit low aleatoric uncertainty, while misclassified pixels will display high aleatoric uncertainty. Analyzing misclassification detection is complex because the ground truth varies across different model predictions. Based on these three frames, we cannot see a large performance difference between the various model variants.

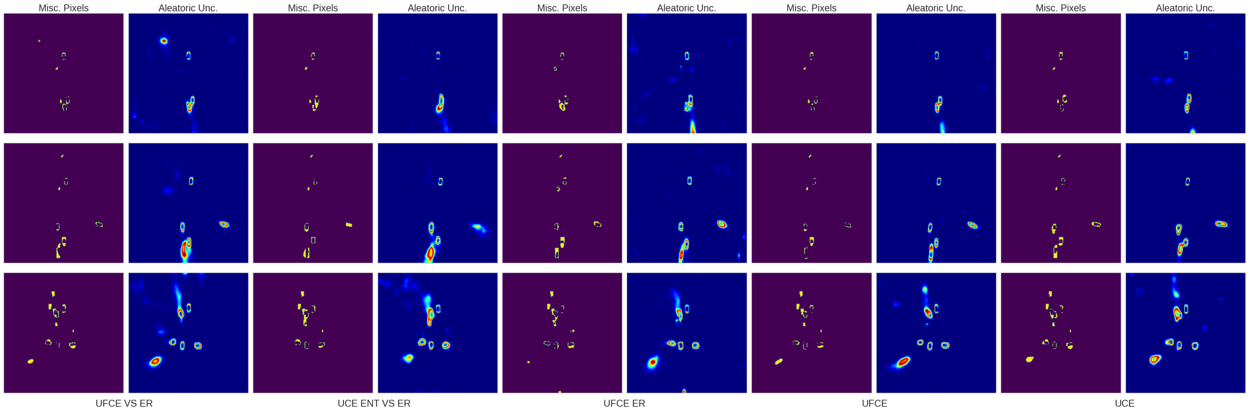

Figure 2: ViZ1

Lastly, we visualize the predicted epistemic uncertainty in Figure 3. We anticipate that in-distribution (ID) pixels exhibit low epistemic uncertainty, and out-of-distribution (OOD) pixels are expected to display high uncertainty levels. It's important to note that our interest primarily lies in the comparative uncertainty levels between ID and OOD pixels rather than the absolute uncertainty scale. Our findings demonstrate that our proposed model, UFCE+ER+VS, achieves the most precise predictions in identifying epistemic uncertainties. We observed a significant increase in false positives upon removing the VS component from the model. When we replace the UFCE with UCE, the ability to delineate OOD boundaries degrades. For UFCE only, it can not effectively identify the OOD pixels.

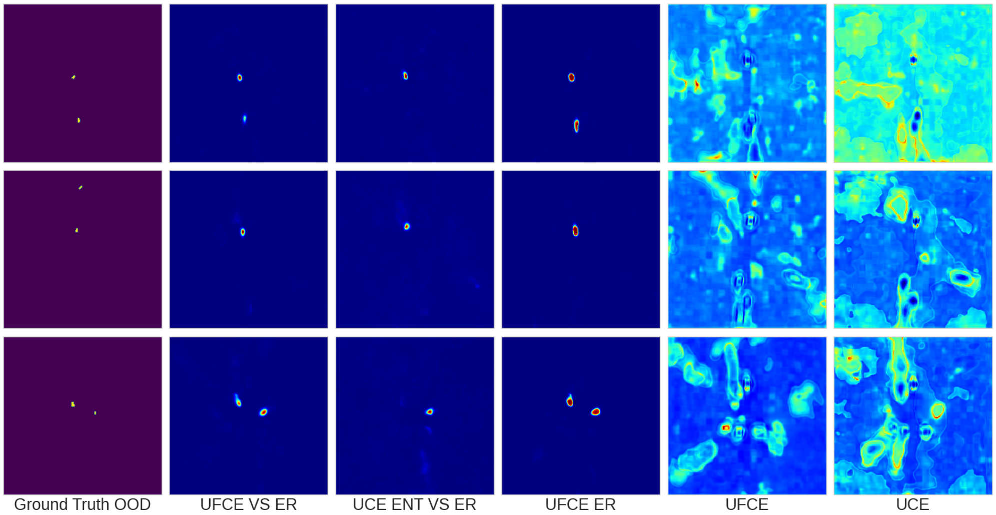

Figure 3: ViZ2

2 Detailed explanations

Why do we pick the two datasets (nuScence and CARLA)? We consider both synthetic and real-world datasets for our experiments. For synthetic data, we utilize the widely recognized CARLA simulator Dosovitskiy et al. (2017) to collect our dataset, which will be made available upon request due to its large size. Our simulated CARLA dataset features five towns with varied layouts and diverse weather conditions to enhance dataset diversity. In terms of real-world data, we employ the nuScenes Caesar et al. (2020) dataset, which is also used in the evaluation of the two segmentation backbones used in our paper: LSS [2] and CVT [3], as well as an updated leaderboard. Based on the introduction of nuScence, KITTI (Behley et al., 2019) primarily features suburban streets with low traffic density and simpler traffic scenarios, with annotations limited to the front camera view instead of a full 360-degree perspective. It also lacks radar data and is designated for non-commercial use only. In contrast, nuScenes aims to enhance these features by providing dense data from both urban and suburban environments in Singapore and Boston.

How do we set the evaluation tasks? In this study, we focus on the problem of uncertainty quantification for BEV semantic segmentation (BEVSS). BEVSS involves predicting the classification of each pixel in the BEV view such as roads, lane marks, vehicles, etc. We aim to improve the quality of uncertainty prediction, assessed through calibration, misclassification detection, and out-of-distribution (OOD) detection while ensuring that the segmentation task's performance remains competitive with existing benchmarks.

Why do we only conduct vehicle and drivable region segmentation? For the segmentation task, we adhere to the experiment settings outlined in literature sources Philion and Fidler (2020) and (Zhou and Krähenbühl, 2022). These sources guide our evaluation of models on object-based segmentation tasks, such as segmenting vehicles and identifying drivable regions.

Which metrics do we use to evaluate the estimated uncertainty? To evaluate the quality of uncertainty quantification, ECE score is for calibration and AUROC/AUPR are for misclassification detection and OOD detection. Specifically, we use the 'animal' class as the OOD for the CARLA dataset, and we use 'motorcycle' as the OOD class for nuScence dataset.

Why are the AUROC and AUPR not consistent? AUROC and AUPR are different ways of measuring the quality of a ranking on data points for separating positives and negatives. A lower AUROC but a higher AUPR for our method indicates that our method produces more true positives among the top-ranked nodes than the evidential UCE model, while the Evidential UCE model can separate true positives and negatives better than our method among the lower-ranked nodes. We note that a high AUPR is important for applications where human experts want to manually verify the top-ranked misclassified nodes or anomalies. Due to the high cost of manual verification, it is important to ensure a high rate of true positives among the top-ranked data points, and in this case, AUPR may be a better metric than AUROC.

Our contribution: First, BEV semantic segmentation models being able to anticipate segmentation errors is critical for reliable application in autonomous driving but under-studied. Based on our knowledge, we are the first work to delve into uncertainty quantification methods for the Bird's Eye View (BEV) semantic segmentation task and build a benchmark. Among the discussed uncertainty quantification models, evidential architecture has better explainability with subjective logic and the inherent ability to disentangle the aleatoric and epistemic uncertainty, coupled with computational efficiency.

Then, we theoretically demonstrated the limitations of prevalent learning function in evidential-based uncertainty quantification models, i.e., UCE. It aims to directly predict the distribution of categorical distribution, i.e. Dirichlet distribution to form opinions in evidential theory. In detail, minimizing UCE loss encourages the predicted Dirichlet distribution to concentrate at the point of a one-hot categorical distribution. This concentration can lead to unreliable predictions of uncertainty.

Next, we propose the new uncertainty focal loss, which can enhance calibration capabilities and address issues caused by imbalanced training data compared to UCE. This is achieved by directing model optimization towards 'hard' examples that exhibit high aleatoric uncertainty, thereby improving both calibration and misclassification detection, as well as the semantic segmentation performance. When comparing our Evidential-UFCE with the Evidential-UCE model, ours consistently demonstrates improved experimental performance, evidenced by higher IoU, lower ECE scores, higher AUROC, and better PR metrics in all twelve configurations.

Lastly, we observe that all models show extremely poor OOD detection performance and then we introduce a strategy of pseudo-OOD exposure to further improve OOD detection performance, which is applied to popular energy baseline and our evidential model. We propose a vacuity scaling approach that focuses model optimization on 'hard' examples characterized by high epistemic uncertainty, aiming to enhance OOD detection. Our evidential-UFCE+vac achieves the best AUROC and AUPR over three baselines in all four configurations, except that AUROC shows 4 percent lower than that of UCE on the nuScenes dataset, but it is over 11 percent higher on the PR.

References

J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall. 2019. SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. In *Proc. of the IEEE/CVF International Conf. on Computer Vision (ICCV)*.

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. In *CVPR*.

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.

R. Kesten, M. Usman, J. Houston, T. Pandya, K. Nadhamuni, A. Ferreira, M. Yuan, B. Low, A. Jain, P. Ondruska, S. Omari, S. Shah, A. Kulkarni, A. Kazakova, C. Tao, L. Platinsky, W. Jiang, and V. Shet. 2019. Lyft Level 5 AV Dataset 2019. `https://level5.lyft.com/dataset/`.

Jonah Philion and Sanja Fidler. 2020. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*. Springer, 194–210.

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.