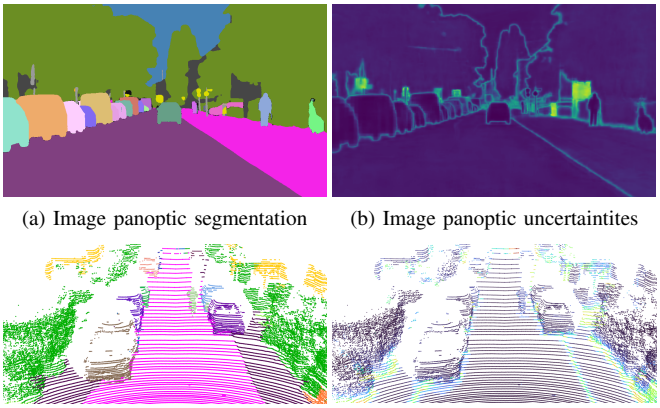
Uncertainty-aware Panoptic Segmentation for Camera and LiDAR Data

Kshitij Sirohi[1], Sajad Marvi[1] and Wolfram Burgard[2]

***Abstract*— Reliable scene understanding is indispensable for modern autonomous systems. Current learning-based methods typically try to maximize their performance based on segmentation metrics that only consider the quality of the segmentation. However, for the safe operation of a system in the real world it is crucial to consider the uncertainty in the prediction as well. In this work, we discuss the task of uncertainty-aware panoptic segmentation, which aims to predict per-pixel semantic and instance segmentations, together with per-pixel uncertainty estimates. We present two novel Evidential Panoptic Segmentation Networks, EvPSNet for solving this task with camera images, and EvLPSNet for LiDAR data. We provide several strong baselines combining state-of-the-art panoptic segmentation networks with sampling-free uncertainty estimation techniques for comparison. Extensive evaluations show that our approaches achieve the new state-of-the-art for the uncertainty-aware Panoptic Quality (uPQ) and the panoptic Expected Calibration Error (pECE). We make our code available at: `https://github.com/kshitij3112`**



(a) Image panoptic segmentation (b) Image panoptic uncertaintites

(c) LiDAR panoptic segmentation (d) LiDAR panoptic uncertainties

Fig. 1: Panoptic segmentations with their associated uncertainties as predicted by our EvPSNet for Cityscapes and EvLPSNet for SemanticKITTI data.

I. Introduction

Due to the recent advances in deep learning, perception systems of modern autonomous systems largely rely on convolutional neural networks (CNNs), in particular for the tasks of semantic segmentation [1] and object detection [2]. However, these two similar tasks are still often treated separately. Aiming for a holistic scene understanding, Kirillov *et al.* [3] introduced panoptic segmentation for combined segmentation of *stuff* classes, consisting of amorphous regions like road surfaces, and *thing* classes, consisting of distinct instances of objects like cars and pedestrians.

Several existing panoptic segmentation methods provide CNN-based architectures for different modalities, such as cameras and LiDARs [4], [5]. Being supervised learning-based approaches, these networks first learn on a training dataset, and then evaluate their performance using specific metrics on a test set. A typical limitation is that the training and test sets usually have quite similar data distributions and conditions, while these can be quite different during deployment in the real world, e.g. due to different weather conditions or unseen objects. Since current training metrics typically consider only the performance, the resulting networks can be quite overconfident in their (false) predictions, possibly posing safety-critical threats in autonomous driving scenarios. In general, these approaches lack insight into the performance of the network in unseen environments and into the reliability of it's confidence estimate output.

In this work, we discuss the novel task of uncertainty-aware panoptic segmentation, illustrated in Fig. 1, intending to provide reliable predictions even in challenging scenarios and to motivate future research in the field of holistic scene understanding. Conventional methods utilize the simple softmax operation to provide probability estimates, which are quite limited in their reliability and typically inflated [6]. On the other hand, sampling-based methods such as dropout are primary candidates employed for reliable uncertainty estimation in various tasks. However, these approaches are computationally intense and thus are not suitable for real-time applications, such as autonomous driving. In this context the research in sampling-free methods for uncertainty estimation is gaining interest. One such method is evidential deep learning, which is already being used successfully in classification [6], regression [7], and multitask learning settings [8].

We propose two novel evidential panoptic segmentation networks, the EvPSNet and EvLPSNet architectures for Camera and LiDAR data, respectively. Applying evidential deep learning, these networks are able to simultaneously predict semantic and instance segmentation outputs and the corresponding pixel-wise or point-wise uncertainties. The predicted uncertainties can be utilized in downstream tasks in a probabilistic manner, helping with a robust and safe performance e.g. of localization algorithms.

II. Related Work

A. Panoptic Segmentation

Since the inception of the panoptic segmentation task, methods have generally taken one of two approaches: either proposal-free (bottom-up) or proposal-based (top-down). In

[1]Department of Computer Science, University of Freiburg, Germany. [2]Department of Engineering, Technical University Nürnberg, Germany. This work was financed by the Baden-Württemberg Stiftung gGmbH.

the proposal-free approach [9], [10] first a semantic segmentation is performed, followed by clustering the pixels belonging to the thing classes. The corresponding clustering methods include center and offset regression [9], calculation of pixels affinity [11], or a Hough voting scheme [12]. In contrast, proposal-based methods [13], [14], [4] consist of two parallel heads, one to perform semantic segmentation and the other to predict bounding boxes and instance masks for thing class objects. One way to learn the panoptic segmentation based on the semantic and instance logits in a parameter-free fashion is using a post-processing fusion technique [4], [15], which, however, does not provide an estimate on the prediction uncertainty. Moreover, the proposed extensions to the panoptic segmentation task also do not provide any uncertainty estimation [16], [17], [18].

Panoptic segmentation of LiDAR data can be done proposal-based as well, e.g. in the case of EfficientLPS [5]. Proposal-free approaches, such as Panoptic-PolarNet [19] utilizes a Panoptic Deeplab-based [9] instance head to regress offsets and centers for different instances. DS-Net [20] proposes a dynamic shifting module to move instance points towards their respective center. Panoptic-PHNet [21] utilizes two different encoders, BEV and voxel-based, to encode point cloud features, followed by a KNN-transformer module to model interaction among voxels belonging to thing classes.

B. Uncertainty Estimation

Uncertainty estimation with neural networks has been popular for quite some time. Bayesian neural networks (BNNs) learn the distribution over network weights to provide a probabilistic model for a network's output. Gal *et al.* [22] proposed the method Monte Carlo (MC) dropout using dropout for variational inference. The disadvantage of sampling-based methods is that they are not fit for real-time applications, as they either need multiple passes through the network or multiple ensemble networks utilizing more computation resources.

Guo *et al.* [23] proposed a sampling-free method called temperature scaling (TS), to learn a scaling factor for the learned logits, calibrating the predicted probabilities. The one disadvantage is that the scaling factor is typically learned on the validation set and thus a bias can be expected. Sensoy *et al.* [6] utilized evidential theory to introduce deep evidential learning to quantify classification uncertainty in a sampling-free fashion. Here, the network is trained to collect parameters for a high-order distribution, the Dirichlet distribution in their case, from which the uncertainty of the prediction is computed.

III. Technical Approach

A. EvPSNet Architecture

We utilize a state-of-the-art proposal-based panoptic segmentation network, EfficientPS [4], as our base network for our EvPSNet architecture, as illustrated in Fig. 2. It consists of a shared EfficientNet [24] backbone with a two-way Feature Pyramid Network (FPN) and separate semantic

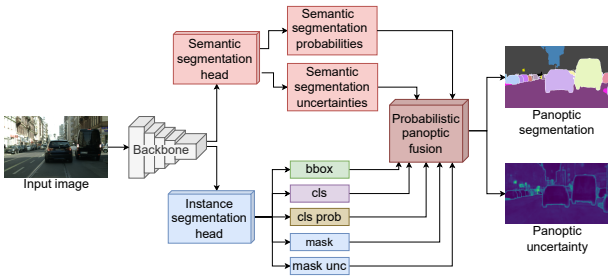


Fig. 2: Overview of our EvPSNet architecture.

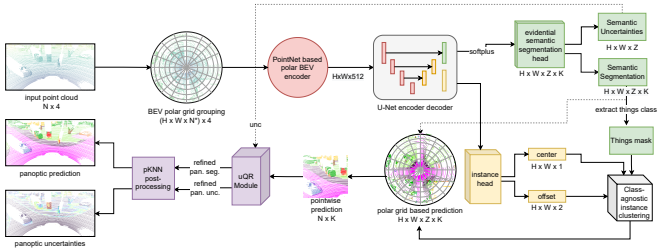


Fig. 3: Overview of our EvLPSnetwork architecture.

and instance segmentation heads. The semantic segmentation head consists of modified Dense Prediction Cell (DPC) [25] modules to capture the contextual information of features at multiple scales. The instance head is a variation of Mask RCNN [26]. As EfficientPS provides only panoptic segmentation, its segmentation heads lack the capability to estimate the uncertainty of its predictions. We employ evidential deep learning [6] to quantify semantic segmentation, instance segmentation, and classification uncertainty in our uncertainty-aware semantic and instance segmentation heads.

We modify the semantic head output by replacing the softmax layer with the softplus activation function. We use the Dirichlet distribution as prior for our per-pixel multinomial classification, parametrized by $\alpha = [\alpha^1, ..., \alpha^C]$, where C is the number of classes and $\alpha_i^c = \text{softplus}(l_i^c) + 1$ for network output logit l for pixel i and class c. The corresponding probability p and uncertainty u are calculated as:

$$p_i^c = \alpha_i^c / S_i \tag{1}$$

$$u_i = C / S_i, \tag{2}$$

where $S_i = \sum_{c=1}^{C} \alpha_i^c$.

The instance head is a modified version of Mask RCNN [26] similar to the EfficientPS architecture. The head consists of a Region Proposal Network (RPN) to generate proposals and objectness scores. The ROI align extracts the features bounded within the generated proposals from the RPN. These features are fed to the separate bounding box regression, classification, and mask generation heads. We focus on providing the uncertainty-aware mask segmentation and object classification. Our novel probabilistic fusion module leverages the predicted probabilities and uncertainties of our semantic and instance segmentation heads to fuse them in a efficient and straight forward way.

B. EvLPSNet Architecture

An overview of our EvLPSNet architecture is shown in Fig. 3. It is based on the proposal-free Panoptic-PolarNet network [19]. Our evidential semantic segmentation head and Panoptic-Deeplab based [9] instance segmentation head utilize the learned features to predict per-point semantic segmentation, semantic uncertainty, instance center and offsets. The predictions from both heads are fused to provide panoptic segmentation results. Leveraging the segmentation uncertainties, our proposed query and refine module helps to improve the prediction for points within uncertain voxels. Moreover, post-processing using our efficient probability-based KNN improves the results further.

We project the LiDAR points into a polar BEV grid utilizing the encoder design proposed by PolarNet [27]. The subsequent encoder-decoder network utilizes the U-net [28] architecture.

We utilize evidential deep learning [6] to provide voxel-level semantic segmentation with calibrated uncertainty estimation, similar to the EvPSNet architecture above. Our uncertainty-based query and refinement module (uQR) leverages the predicted uncertainties to counter the discretization errors due to the BEV grid structure. We select the top 20k most uncertain points and pass them to our uQR module to actively improve the segmentation quality in an efficient way. Further, our probability-based k nearest neighbors (pKNN) approach aims for efficient refining by considering only points that have a probability (Eq. (1)) below a certain threshold, followed by a majority voting to decide the final label.

C. Metrics for uncertainty-aware panoptic segmentation

a) Calibration metric: Here, we aim to provide a metric capable of evaluating the network calibration, i.e. how well the predicted confidence matches the actual accuracy the prediction, for the panoptic segmentation task. A common measure for the calibration accuracy is the Expected Calibration Error (ECE) [29]. However, as pointed out by Nixon *et al.* [30], the ECE has some severe limitations, in particular: First, it takes only the maximum class probability of the prediction into account, ignoring the probability of all other classes. Second, it is only suited to evaluate the calibration of the semantic segmentation task, while the instance segmentation is ignored.

To solve the first limitation, we propose to employ the predicted uncertainty, rather than just the highest class probability. Considering $u_i \in [0, 1]$ to be the predicted uncertainty for pixel i, we define the corresponding confidence as $\text{conf}_i = 1 - u_i$. Further, we define the corresponding accuracy as $\text{acc}_i = 1$ if the predicted class matches the ground truth, and $\text{acc}_i = 0$ otherwise. For each image we partition conf_i into B equally spaced bins and calculate the average confidence $\text{conf}(b)$ and average accuracy $\text{acc}(b)$ for each bin b. Then we define a novel uncertainty-aware calibration metric as

$$\text{uECE} = \sum_{b=1}^{B} \frac{|b|}{N} \left|\text{acc}(b) - \text{conf}(b)\right|, \tag{3}$$

where $|b|$ is the number of pixels in bin b, and N is the total number of pixels. This definition is analogous to the original ECE, but using the confidence instead of the highest probability output, solving the first limitation.

To solve the second limitation, we proceed to additionally incorporate the instances of the scene into a metric. The first step is to identify correctly predicted segments f, which is done by selecting those that have $\text{IoU} > 0.5$ with a ground truth segment g. Here, a segment can either be the mask of a single instance for *thing* classes, or all pixels belonging to a *stuff* class. This leads to M unique matching pairs (f, g) [3]. We then calculate uECE for each of these correctly predicted segment separately, and take the average to form our novel panoptic calibration metric:

$$\text{pECE} = \frac{1}{M} \sum_{(f,g)} \text{uECE}(f, g). \tag{4}$$

Due to the matching, this quantity is sensitive to the calibration within the instances of *thing* classes, as well as to the *stuff* classes.

b) Overall performance metric: Our next aim is to provide a unified metric to evaluate the panoptic segmentation and uncertainty prediction together. It is based on the common Panoptic Quality (PQ):

$$\text{PQ} = \frac{\sum_{(f,g)} \text{IoU}(f, g)}{\text{TP} + \frac{1}{2}\text{FP} + \frac{1}{2}\text{FN}} \tag{5}$$

where the IoU is calculated for the matched pairs (f, g), and TP, FP and FN are the number of true positive, false positive, and false negative segments, respectively. We define our novel uncertainty-aware panoptic quality uPQ by combining pECE with PQ as:

$$\text{uPQ} = (1 - \text{pECE})\text{PQ} \tag{6}$$

IV. Experimental Evaluation

We evaluate the performance of our networks on the Cityscapes dataset [31], providing Camera images for EvPSNet, and the SemanticKITTI [32] dataset, providing LiDAR data for EvLPSNet.

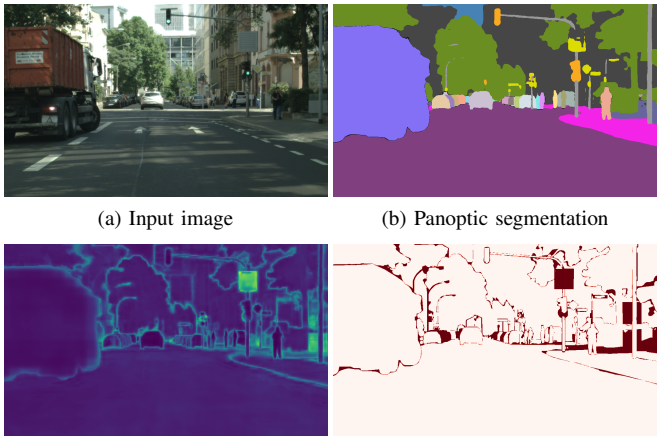
We provide several baselines, first the state-of-the-art top-down and bottom-up networks, EfficientPS [4] and PanopticDeepLab [9], for the panoptic segmentation of images without uncertainty-awareness. For the LiDAR data, we choose the proposal-based EfficientLPS [5] and the proposal-free Panoptic-PolarNet [19]. To extract uncertainties, we chose the sampling-free methods temperature scaling (TS) and evidential learning (Ev). We estimate the uncertainty output for the original networks and their TS variants by calculating the normalized entropy of the predicted probabilities. The evidential variant replaces the softmax with the softplus operation and training the semantic segmentation head with the evidential loss.

TABLE I: Performance values in % on the Cityscapes validation set. Lower values are better for ↓, and larger values otherwise.

Method	uPQ	PQ	SQ	RQ	uPQ^{Th}	PQ^{Th}	SQ^{Th}	RQ^{Th}	uPQ^{St}	PQ^{St}	SQ^{St}	RQ^{St}	pECE ↓
EfficientPS [4]	49.9	63.5	81.6	76.9	47.1	**58.9**	80.7	**72.8**	52.0	66.8	82.2	79.8	21.3
EfficientPS [4] + TS [23]	50.1	63.5	81.6	76.9	44.6	58.8	80.6	**72.8**	54.2	66.8	82.2	79.8	21.1
PDeepLab [9]	47.7	63.1	82.0	75.9	42.6	55.7	80.6	68.9	51.45	68.5	82.9	81.0	24.3
PDeepLab [9] + TS [23]	50.5	63.1	82.0	75.9	43.5	55.7	80.6	68.9	56.8	68.5	82.9	81.0	20.5
PDeepLab [9] + Ev	51.6	62.5	**82.1**	75.1	43.5	54.3	**80.9**	66.9	57.7	68.5	**83.0**	81.0	17.5
EvPSNet (ours)	**54.9**	**64.1**	81.4	**77.8**	**50.4**	56.8	80.2	70.9	**57.9**	**69.5**	82.3	**82.8**	**14.3**

TABLE II: Performance values in % on the SemanticKITTI validation set. Lower values are better for ↓, and larger values otherwise.

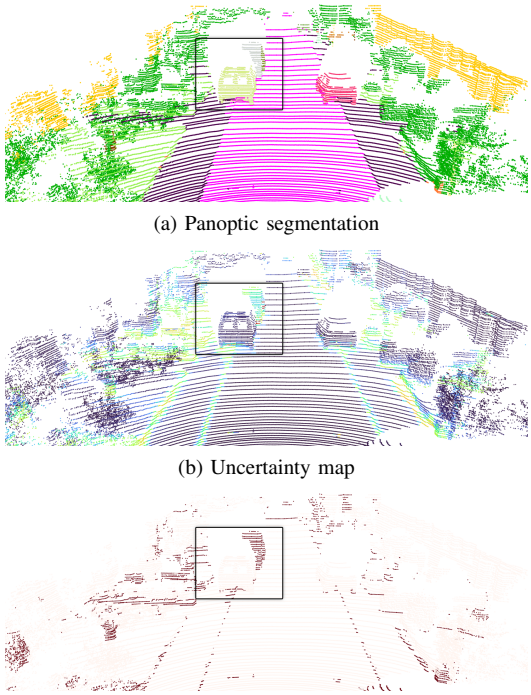
Method	uPQ	PQ	pECE ↓	uPQ^{Th}	PQ^{Th}	$pECE^{Th}$ ↓	uPQ^{St}	PQ^{St}	$pECE^{St}$ ↓	uECE ↓	mIOU
EfficientLPS	48.7	57.1	14.6	52.6	59.9	**12.1**	45.9	**54.9**	16.4	16.5	62.3
EfficientLPS + TS	49.6	56.9	12.9	48.5	59.8	18.8	50.2	54.9	8.7	7.7	62.1
Panoptic-PolarNet	48.8	**58.5**	16.6	**53.1**	**65.7**	19.1	45.4	53.3	14.7	13.2	63.2
Panoptic-PolarNet + TS	48.1	**58.5**	17.7	51.4	65.7	21.8	45.4	53.3	14.7	10.5	63.3
EvLPSNet	**51.4**	58.0	**11.5**	52.7	62.7	15.9	**50.1**	54.6	**8.2**	**7.1**	**64.0**



(a) Input image (b) Panoptic segmentation

(c) Uncertainty map (d) Error map

Fig. 4: Qualitative results of EvPSNet on Cityscapes data.



(a) Panoptic segmentation

(b) Uncertainty map

(c) Error map

Fig. 5: Qualitative results of EvLPSNet on SemanticKITTI data.

To compare the performance we employ our proposed uncertainty-aware panoptic segmentation metrics, as well as standard panoptic segmentation metrics. The results for EvPSNet are presented in Tab. I and for EvLPSNet in Tab. II. Our methods outperform all baselines on the uncertainty-aware panoptic segmentation metrics uPQ and pECE.

We further provide qualitative results, including the panoptic segmentation, uncertainty and error maps in Fig. 4 for EvPSNet and in Fig. 5 for EvLPSNet. Comparing the predicted uncertainty with the error maps a high correlation can be observed, which provides clear visual validation of the approaches.

V. Conclusions

In this work, we discussed the novel task of uncertainty-aware panoptic segmentation. To this end, we provided two metrics, uPQ and pECE, for evaluating and comparing the performance on our proposed task. We also introduced several strong baselines by combining state-of-the-art panoptic segmentation networks with sampling-free uncertainty estimation techniques. Our proposed EvPSNet and EvLPSNet architectures outperforms all the baselines on the uncertainty-aware panoptic metrics.

References

[1] J. Vertens, J. Zürn, and W. Burgard, "Heatnet: Bridging the day-night domain gap in semantic segmentation with thermal images," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 8461–8468.

[2] J. Zürn and W. Burgard, "Self-supervised moving vehicle detection from audio-visual cues," *arXiv preprint arXiv:2201.12771*, 2022.

[3] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, "Panoptic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9404–9413.

[4] R. Mohan and A. Valada, "Efficientps: Efficient panoptic segmentation," *International Journal of Computer Vision*, vol. 129, no. 5, pp. 1551–1579, 2021.

[5] K. Sirohi, R. Mohan, D. Büscher, W. Burgard, and A. Valada, "Efficientlps: Efficient lidar panoptic segmentation," *IEEE Transactions on Robotics*, 2021.

[6] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[7] A. Amini, W. Schwarting, A. Soleimany, and D. Rus, "Deep evidential regression," *Advances in Neural Information Processing Systems*, vol. 33, pp. 14 927–14 937, 2020.

[8] K. Petek, K. Sirohi, D. Büscher, and W. Burgard, "Robust monocular localization in sparse hd maps leveraging multi-task uncertainty estimation," *arXiv preprint arXiv:2110.10563*, 2021.

[9] B. Cheng, M. D. Collins, Y. Zhu, T. Liu, T. S. Huang, H. Adam, and L.-C. Chen, "Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 475–12 485.

[10] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, and L.-C. Chen, "Axial-deeplab: Stand-alone axial-attention for panoptic segmentation," in *European Conference on Computer Vision*. Springer, 2020, pp. 108–126.

[11] N. Gao, Y. Shan, Y. Wang, X. Zhao, Y. Yu, M. Yang, and K. Huang, "Ssap: Single-shot instance segmentation with affinity pyramid," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 642–651.

[12] B. Leibe, A. Leonardis, and B. Schiele, "Combined object categorization and segmentation with an implicit shape model," in *Workshop on statistical learning in computer vision, ECCV*, vol. 2, no. 5, 2004, p. 7.

[13] A. Kirillov, R. Girshick, K. He, and P. Dollár, "Panoptic feature pyramid networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6399–6408.

[14] Y. Li, X. Chen, Z. Zhu, L. Xie, G. Huang, D. Du, and X. Wang, "Attention-guided unified network for panoptic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7026–7035.

[15] Y. Xiong, R. Liao, H. Zhao, R. Hu, M. Bai, E. Yumer, and R. Urtasun, "Upsnet: A unified panoptic segmentation network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8818–8826.

[16] R. Mohan and A. Valada, "Amodal panoptic segmentation," *arXiv preprint arXiv:2202.11542*, 2022.

[17] N. Gosala and A. Valada, "Bird's-eye-view panoptic segmentation using monocular frontal view images," *IEEE Robotics and Automation Letters*, 2022.

[18] D. de Geus, P. Meletis, C. Lu, X. Wen, and G. Dubbelman, "Part-aware panoptic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5485–5494.

[19] Z. Zhou, Y. Zhang, and H. Foroosh, "Panoptic-polarnet: Proposal-free lidar point cloud panoptic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 194–13 203.

[20] Y. Zhao, X. Zhang, and X. Huang, "A divide-and-merge point cloud clustering algorithm for lidar panoptic segmentation," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 7029–7035.

[21] J. Li, X. He, Y. Wen, Y. Gao, X. Cheng, and D. Zhang, "Panoptic-phnet: Towards real-time and high-precision lidar panoptic segmentation via clustering pseudo heatmap," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 809–11 818.

[22] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*. PMLR, 2016, pp. 1050–1059.

[23] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1321–1330.

[24] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[25] L.-C. Chen, M. Collins, Y. Zhu, G. Papandreou, B. Zoph, F. Schroff, H. Adam, and J. Shlens, "Searching for efficient multi-scale architectures for dense image prediction," *Advances in neural information processing systems*, vol. 31, 2018.

[26] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[27] Y. Zhang, Z. Zhou, P. David, X. Yue, Z. Xi, B. Gong, and H. Foroosh, "Polarnet: An improved grid representation for online lidar point clouds semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9601–9610.

[28] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[29] M. P. Naeini, G. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using bayesian binning," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[30] J. Nixon, M. W. Dusenberry, L. Zhang, G. Jerfel, and D. Tran, "Measuring calibration in deep learning." in *CVPR Workshops*, vol. 2, no. 7, 2019.

[31] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[32] J. Behley, A. Milioto, and C. Stachniss, "A benchmark for lidar-based panoptic segmentation based on kitti," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 596–13 603.