2024 IEEE Intelligent Vehicles Symposium (IV)
June 2-5, 2024. Jeju Island, Korea

Obstacle-sensitive Semantic Bird-Eye-View Map Generation with Boundary-aware Loss for Autonomous driving

Shuang Gao, Qiang Wang, and Yuxiang Sun

2024 IEEE Intelligent Vehicle Symposium (IV) | 979-8-3503-4881-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/IV55156.2024.10588473

Abstract— **Detection of road obstacles is important for autonomous driving. However, road obstacles, like pedestrians, usually account for quite a small portion compared with other semantics, such as road layouts. This leads to the class-imbalance problem in real-world driving datasets and hinders environment perception for autonomous driving. In this paper, we propose an obstacle-sensitive network to improve the semantic Bird-Eye-View (BEV) map generation performance for minority classes. To this end, a context-depth attention module and a boundary-aware loss are introduced. We conduct ablation studies to verify the effectiveness of the proposed network. We also compare our network with other semantic BEV map generation methods. The results demonstrate that our network achieves better performance in terms of semantic BEV map generation, especially for minority classes.**

I. Introduction

An efficient data representation for perception of surrounding traffic environments for autonomous vehicles is necessary. The semantic bird-eye-view (BEV) map is a type of popular data representation due to its efficiency. It is easy to fuse different information from multi-modal inputs, such as visual images and LiDAR point clouds, under BEV. Moreover, semantic segmentation provides the abstract information of the surrounding environment, which can bridge the gap between the real world and the simulation environment, and is also an effective tool to provide fundamental information for downstream tasks, such as trajectory prediction [1]–[4] and autonomous navigation [5]–[8]. In addition, the BEV map is more flexible in representing dynamically-changing environments. It eliminates the visual differences in the scale of the same object caused by distances. The semantic BEV map generation has gradually attracted great attention in the autonomous driving community.

Detecting obstacles on roads, such as traffic cones and pedestrians, is critical for safe navigation. Correctly detecting those objects could provide early warning signals for decision-making [9]–[11]. However, due to the nature of the practical driving environments, obstacles on roads account for a relatively small portion compared to road layouts. For deep-learning-based autonomous driving, the training data collected in real driving scenarios is also facing such class-imbalanced challenge, leading to degraded performance in segmenting road obstacles. Most existing semantic BEV map generation methods focus on the design of network structures and the improvement of the overall segmentation performance across all classes. They ignore the imbalanced data distribution. In MonoLayout [12], the authors used two branches to segment static and dynamic objects separately. But for dynamic objects, they only predict vehicles, leaving smaller categories out. To improve the segmentation accuracy, we pay more attention to small objects, which occupies a small fraction of the class-imbalanced dataset.

To address the class-imbalanced problem, some work [13]–[15] attempt to use different data augmentation methods to increase the data diversity. The oversampling technique [16]–[18] intends to balance the majority and minority classes by generating new minor data. Such data-level methods are independent of the algorithms. Cost-sensitive approaches [19], [20] solve this problem in another way. Those methods pay attention to designing loss functions rather than the input data.

Unlike the previous works, we present an obstacle-sensitive network containing a context-depth attention module and a boundary-aware loss to handle the class-imbalanced problem for semantic BEV map generation. The context-depth attention module stresses the interaction between the front-view feature maps and the depth distribution. The boundary-aware loss further improves the performance of small object segmentation by emphasizing the margin between different semantic classes. To the best of our knowledge, this is the first work exploring the class-imbalanced problem in the semantic BEV map generation task. The contributions of work are summarized as follows:

1) We design a context-depth attention module to extract the associated features from the front view feature and depth contribution.
2) We propose a novel boundary-aware loss to balance the learning weight of the minority classes.
3) We demonstrate the superiority of our obstacle-sensitive network by comparing it with state-of-the-art methods via extensive experiments.

The rest of this paper is organized as follows: Section II provides a review of related work. Section III explains the details of our network. Section IV presents the experimental results. The last section concludes our work and explores potential avenues for future research.

This work was supported in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2022A1515010116, in part by Hong Kong Research Grants Council under Grant 15222523, and in part by City University of Hong Kong under Grant 9610675. *(Corresponding author: Yuxiang Sun.)*

Shuang Gao is with The Hong Kong Polytechnic University, Kowloon, Hong Kong, and also with Harbin Institute of Technology, Harbin, China (email: gaoshuang.sarah@outlook.com).

Qiang Wang is with Harbin Institute of Technology, Harbin, China (email: wangqiang@hit.edu.cn).

Yuxiang Sun is with City University of Hong Kong, Kowloon, Hong Kong (e-mail: yx.sun@cityu.edu.hk, sun.yuxiang@outlook.com).

II. Related Work

A. Semantic Segmentation Networks

Semantic segmentation aims to assign a semantic label to each pixel in an image [21]–[23]. With the emergence of deep-learning methods, semantic segmentation has become one of the core tasks in computer vision. Unlike classification, the space structure of the image should be preserved in semantic segmentation. In theory, a bigger receptive field could lead to better performance. In 2017, Zhao *et al.* [24] proposed PSPNet with the Pyramid Pooling Module (PPM), which generates the multi-scale feature. This pyramid structure has inspired many follow-up work. DeepLab V2 [25] introduced the Atrous Spatial Pyramid Pooling (ASPP) module, using the atrous convolution to replace the conventional convolution. Later, the same authors [26] proposed DeepLab V3+ with a densely connected encoder-decoder structure. Fu [27] using the stacked deconvolution layers to upsample the feature maps in their proposed SDNet. The Vision Transformer methods have recently provided a new insight into semantic segmentation. TransUnet [28] and Swin-Unet [29] combine the Transformer structure and the well-known segmentation network, U-Net.

B. Semantic BEV Map Generation

Different from the typical semantic segmentation, the semantic BEV map generation requires performing the view transformation and semantic prediction simultaneously. It is more complex but useful than the front-view semantic segmentation for autonomous driving tasks. VED [30] is the pioneer in semantic BEV map prediction. The authors segment the BEV space with a convolutional variational encoder-decoder structure. Mani [12] proposed Monolayout, in which the adversarial feature is leveraged to predict the occupancy of the BEV grid. Pan *et al.* [31] utilize the domain adaptation technique to train with the real RGB images and the synthetic masks in their View Parsing Network (VPN). S2G2 [32] explored a semi-supervised manner of generating the semantic BEV map. LSS [33] introduces depth prediction into the semantic BEV map generation. Inspired by LSS, BEVDepth [34] improved the depth prediction module with additional depth supervision during the training phase.

C. Class-imbalanced Learning

The deep-learning dataset collected in the real world often has a long-tailed distribution, resulting in insufficient learning of the minority categories. To tackle the class-imbalance problem, many methods are proposed, which can be divided into three main categories: 1) data-level approaches [13]–[15], [17] focus on rebalancing the input data distribution by expanding the data diversity or oversampling the minority data; 2) algorithmic-level approaches [35], [36] attempt to modify the typical classification methods to predict the minority classes better; 3) cost-sensitive approaches [19], [20] propose the new loss functions to guide the network training with the bigger weight assigned to the minority classes.

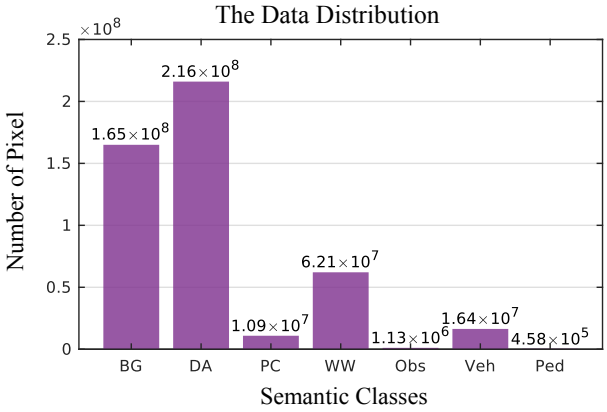

Fig. 1: The data distribution for different semantic classes in the nuScenes dataset. For our semantic BEV map generation task, we create the semantic BEV ground truth for 7 categories, including the background (BG), drivable area (DA), pedestrian crossing (PC), walkway (WW), obstacle (Obs), vehicle (Veh), and pedestrian (Ped).

For autonomous driving tasks, solving the class-imbalanced problem is important for safe driving. This work explores this problem in the semantic BEV map generation. To this end, we propose a depth-context attention module and a boundary-aware loss function to improve the segmentation of the minority semantic classes.

III. The Proposed Method

A. The Overall Architecture

The motivation of this work is to explore the class imbalance problem, which is quite common in autonomous driving datasets. For semantic BEV map generation, we employ the nuScenes [37] dataset as training data for this work. The statistics of the number of pixels in each category are illustrated in Fig. 1, from which we find that the data of the dynamic categories (obstacle, vehicle, and pedestrian) are much less than that of the static ones (background, drivable area, pedestrian crossing, and walkway). Such an imbalance in the data distribution could degrade the performance of semantic BEV map generation. To tackle this problem, we propose an obstacle-sensitive semantic BEV map generation network, shown in Fig. 2.

The proposed network mainly consists of a context-depth attention module and a boundary-aware loss to cope with the data imbalance problem. Our network takes as input the front-view monocular image and the corresponding camera matrix. Using the given camera's intrinsic and extrinsic matrices, the 3D position P_w in the world coordinate of each pixel P_i in the image can be calculated. However, the depth estimation from the monocular image is an ill-posed problem. We follow [33] to set n possible depths $d_1, d_2, ..., d_n$ for the given scene first and get a set of points in the world coordinate. This process is represented by the lift module in Fig. 2. At the same time, the encoder predicts the front-view feature f_{front} and the depth distribution probability $\mathcal{P}_d$

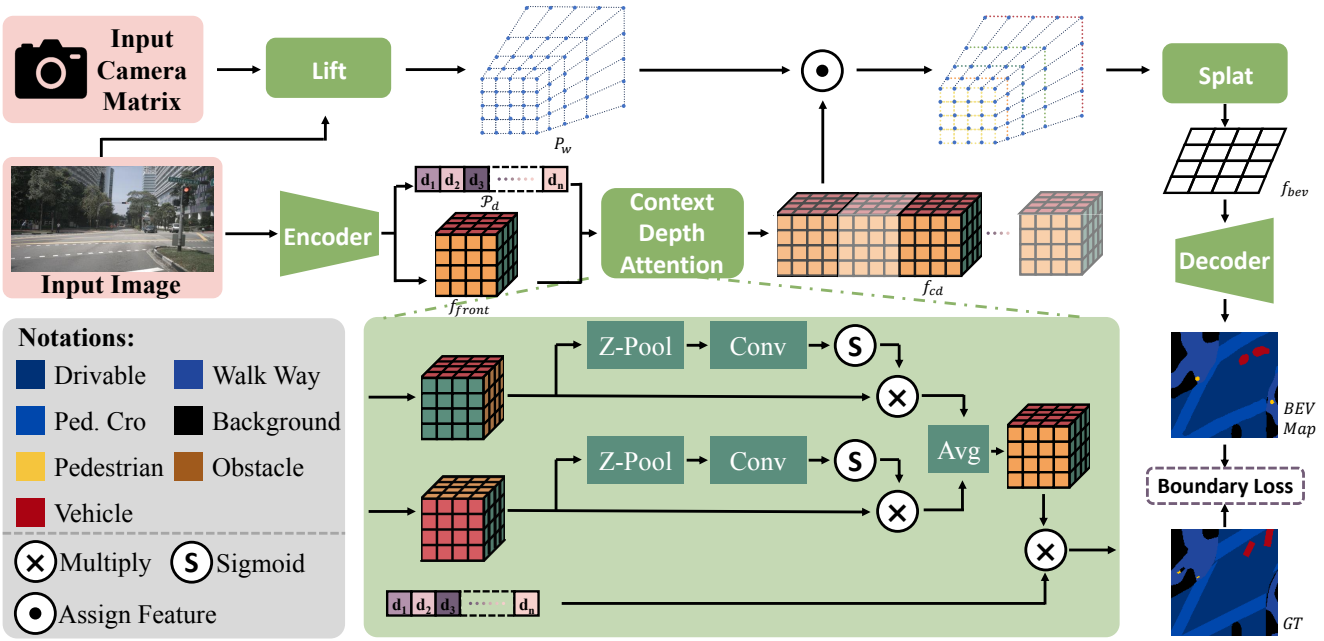

Fig. 2: The overall structure of our proposed network. Our network takes as input the front-view image and the camera intrinsic and extrinsic matrices. Following [33], each pixel in the front-view image is lifted to the 3D world coordinate, forming a set of points. Then, we use the proposed context-depth attention module to get the feature map for each point. After the splatting and decoding operation, the semantic BEV map is generated. To train our network with the class-imbalanced dataset, we design a boundary-aware loss function to emphasize the minority classes during the training phase.

according to the input image. Here, we design a context-depth attention module to extract the salient features f_{cd} for the various depths. Then, f_{cd} is assigned to each projected point. We get the semantic BEV map after a splatting operation and the decoder module. To make the whole network more sensitive to small objects, a boundary-aware loss is used to supervise the network training by assigning a bigger weight to minority semantic classes.

B. The Context-Depth Attention Module

Since we combine the predicted front-view feature map f_{front} and the depth distribution $\mathcal{P}_d$ as the feature tensor for each point in the 3D coordinate, the interaction between the feature maps and the depth distribution should be investigated to assign the point with a more reasonable feature. Therefore, we take advantage of the attention mechanism to explore the relationship between the spatial feature and the depth. In this section, we take inspiration from [38], which extracts the salient feature from various dimensions by rotating the feature tensor. This attention module is modified to adapt to our task for capturing the context-depth dependencies.

The context-depth attention module is shown in the green background of Fig. 2. Following [38], there are three branches in this attention module. The input tensor of this module is the output feature map from the encoder, f_{front}. The shape of the input tensor is $B \times C \times h \times w$, where B, C, h, w are the batch size, channel dimension, and spatial dimensions. The feature maps are respectively rotated along the h or w dimensions before being fed into the first two branches. The rotation operation enables the attention module to better stress on the interaction between the different spatial dimensions and the channel dimension without the loss of the spatial information. The first two branches have the same structure, consisting of a Z-pooling operation and a standard $k \times k$ convolutional layer. After the two operations, the attention score can be obtained. Then, we multiply the input tensor with the score to get the attention-based tensor. The third branch deals with the depth information and the attention-based features by element-wise multiplication. The output tensor of this module embraces the context feature between the different dimensions and the depth distribution, capturing more details from the front-view images.

C. The Boundary-aware Loss

The small road obstacles, such as pedestrians, account for quite a small portion of the dataset, leading to inferior segmentation performance on those small objects. To stress those minority semantic classes, it is intuitive to increase the weight of the road obstacles in the loss calculation during the training phase. Here, we design a boundary-aware loss to cope with the class-imbalanced problem. According to the statistics of the number of semantic pixels among the different classes, we classify the obstacle, vehicle, and pedestrian as the minority class. First, we assemble those objects on one binary mask, shown as the (c) in Fig. 3. Subsequently, the edges of the objects are extracted, and the edge width is denoted as γ ((d) in Fig. 3). The boundary-

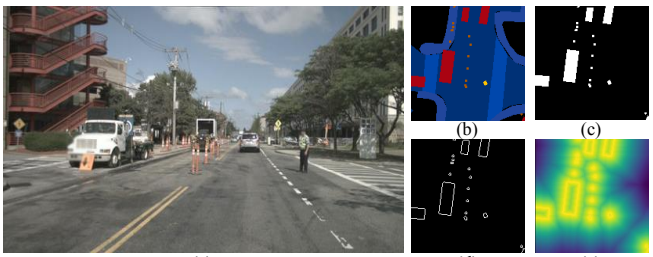

Fig. 3: The visualization of the boundary-aware loss. (*a*) is the input front-view image. (*b*) is the corresponding ground truth. (*c*) is the binary mask for the small objects that appear in the image. (*d*) is the extracted edges of the small objects. (*e*) is the boundary-aware score. Note that the brighter color represents the higher score. The figure is best viewed in color.

aware score S is then generated by calculating the pixel distance from the extracted edges. A small distance will get a higher score. The boundary-aware score is visualized as (*e*) in Fig. 3. Note that the brighter color stands for the higher boundary-aware score. During training, the boundary-aware score is applied to the loss calculation to attach more attention to the small objects as the following formula:

$$L_{BD} = \sum_{i=1}^{n} S_i (Y_i - P_i)^2, \tag{1}$$

where S_i is the boundary-aware score for the pixel i. Y_i and P_i represent the semantic label and the predicted results, respectively.

IV. Experimental Results and Discussions

A. The Dataset and Training Details

We train our proposed network with the public autonomous driving dataset, nuScenes [37]. The nuScenes dataset provides 850 annotated scenes, with 3D object bounding boxes, high-definition (HD) maps, and semantic labels. We project the 3D object bounding boxes into the HD map to obtain the semantic BEV ground truth label. The whole dataset is randomly divided into 548 training sets, 150 validation sets, and 148 testing sets, removing the corrupted data due to the projection.

The network is trained with NVIDIA RTX 3090 GPU. We set the initial learning rate to 1×10^{-4}, and the weight decay is 1×10^{-5}. We adopt EfficientNet-B0 [39] as the backbone to extract the front-view feature. The edge width is a hyper-parameter in the boundary-aware loss, and we set it to 2.0. The selection details are discussed in the ablation study.

B. Ablation Study

1) Ablation on the Network Structure: To make our network sensitive to small objects, such as road obstacles, we design the context-depth attention module and the boundary-aware loss. In this ablation study, we compare the performance of the network variants with or without certain modules.

TABLE I: The results (%) of the ablation study on the network structure. '✓' means that the network includes a certain module. '–' means the module is not contained in the network. We report the performance in terms of majority and minority classes, respectively. The best results are highlighted in bold font.

Variants		Majority		Minority		mIoU	mAP
Att	BL	mIoU	mAP	mIoU	mAP		
–	–	43.22	61.62	8.61	28.14	28.39	47.27
✓	–	46.53	64.66	10.71	28.30	31.18	49.07
–	✓	**49.88**	66.52	9.78	28.28	32.69	50.13
✓	✓	49.70	**66.56**	**11.70**	**28.41**	**32.84**	**50.21**

TABLE II: The results (%) of the ablation study on the hyper-parameter in the boundary-aware loss. γ is the edge width of the small objects. We set 5 values to γ to test the segmentation performance. The results are reported in the majority and minority categories. The best results are highlighted in bold font.

γ	Majority		Minority		mIoU	mAP
	mIoU	mAP	mIoU	mAP		
0.5	43.50	**68.56**	9.23	28.96	32.17	**51.58**
1.0	43.54	65.00	9.86	29.83	32.47	49.92
1.5	49.56	66.74	8.98	29.94	32.62	49.23
2.0	**49.70**	66.56	**11.70**	28.42	**32.84**	50.21
2.5	49.54	65.07	8.05	**31.37**	31.76	50.63

Tab. I shows the results of the segmentation performance of different variants. '✓' represents the variant that includes the certain module, and '–' means not. We divide the 7 semantic classes into majority and minority categories. The former includes background, drivable area, pedestrian crossing, and walkway. The latter consists of obstacle, vehicle, and pedestrian. The table shows that the variant with both context-depth attention module and boundary-aware loss achieves the best performance for small road objects. The results demonstrate the effectiveness of the proposed modules for small object segmentation.

2) Ablation on the Boundary-Aware Loss: In the boundary-aware loss, a hyper-parameter γ controls the edge width for the small objects in the binary mask (shown in the (*d*) of Fig. 3). It influences the calculation of the boundary-aware score. To select the best γ, we conduct the ablation study to compare the performance from the network with different γ. We set 5 values here. The results are displayed in Tab. II. We find that the proposed network performs best when γ is set to 2.0.

C. Comparative Study

We compare our proposed network with the existing semantic BEV map generation networks: PYVA [40], VED [30], LSS [33]. We train those networks with the nuScenes dataset for 20 epochs. The mean Intersection over Union (mIoU) and the mean Average Precision (mAP) are chosen to evaluate the performance of the different networks. This work aims to improve the segmentation performance of

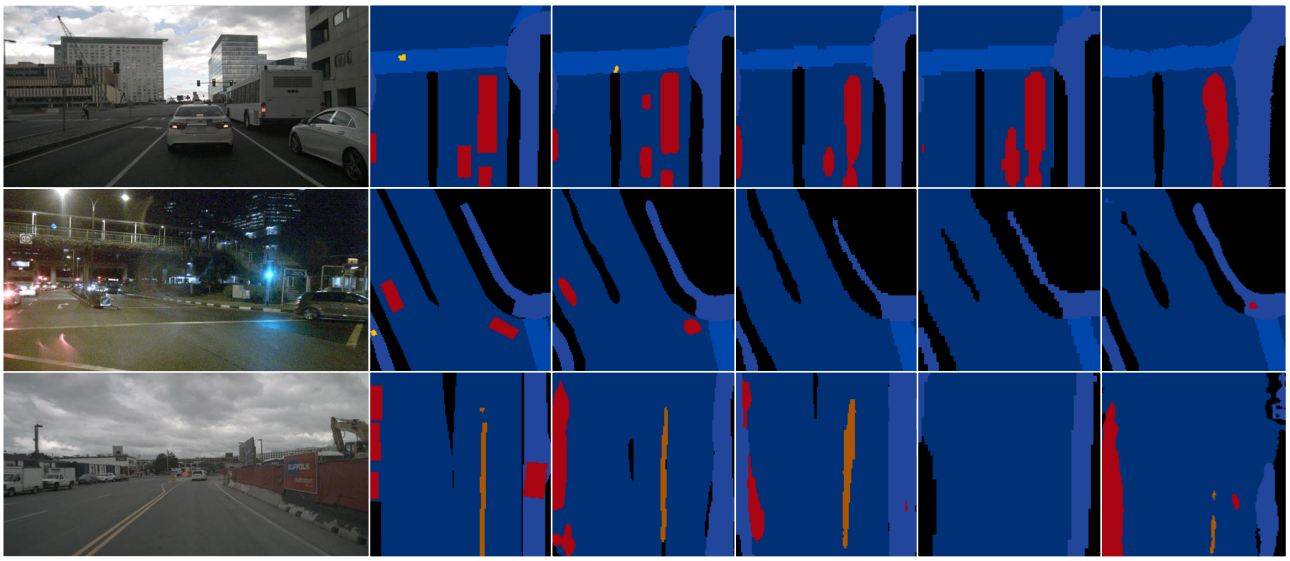

Fig. 4: Sample qualitative demonstrations for the semantic BEV map generation networks. The results demonstrate the superiority of our network. The figure is best viewed in color.

TABLE III: The comparative results (%) compared with the existing semantic BEV map generation networks. As this work aims to improve the segmentation performance of small objects, the results of the minority classes, like obstacle, vehicle, and pedestrian should draw more attention. The best results are highlighted in bold font.

Network	Background		Drivable Area		Ped. Crossing		WalkWay		Obstacle		Vehicle		Pedestrain		mIoU	mAP
	IoU	AP	IoU	AP	IoU	AP	IoU	AP	IoU	AP	IoU	AP	IoU	AP		
PYVA	**63.00**	**73.91**	**70.88**	**84.32**	23.19	44.70	39.73	55.43	8.27	33.84	18.02	39.77	0.00	0.00	31.87	47.42
VED	56.32	71.99	67.36	74.40	27.31	59.74	33.57	**60.22**	0.00	0.00	3.60	**53.80**	0.00	0.00	26.88	45.74
LSS	51.45	66.44	62.87	73.91	27.57	54.13	30.98	52.01	13.64	51.57	11.71	30.31	0.48	2.53	28.39	47.27
OURS	58.44	69.85	67.75	81.96	**36.64**	**60.19**	**35.97**	54.23	**16.12**	**54.57**	**18.21**	27.47	**0.76**	**3.21**	**32.84**	**50.21**

small objects in the driving environment. So, in this comparative experiment, we pay attention to the segmentation results in the minority classes, including obstacle, vehicle, and pedestrian.

1) Quantitative Results: The comparative results of the above-mentioned networks are listed in Tab. III. Our proposed obstacle-sensitive network achieves the best performance in mIoU and mAP over the total 7 classes. The segmentation results of the two majority classes (i.e., background and drivable area) are inferior to that of PYVA, while we have a decided advantage in predicting the semantics for the minority classes (i.e., obstacle, vehicle, and pedestrian), compared with our counterparts. Especially for the pedestrian class, most methods fail to generate the correct segmentation. The results demonstrate superior performance in sensing the small objects in the driving environment.

2) Qualitative Demonstrations: Some sample qualitative demonstrations of the above-mentioned methods are shown in Fig. 4. The figure shows that our obstacle-sensitive network generates the most accurate and clear semantic BEV map. We can see that the objects predicted by our network are close to the real sizes from the first row. The second row shows the semantic BEV map generation at night, from which we find that only our network could infer the positions of the vehicles. The prediction results of the small object (the road obstacles) are shown in the last row. The results illustrate the superiority of our network.

V. Conclusions and Future Work

For environment perception, class imbalance is a common issue in real-world collected datasets, which hinders the correct detection of small objects in driving environments. In this work, we improved the sensing accuracy with the imbalanced data in the form of a semantic BEV map. To this end, we proposed a network composed of a context-depth attention module and a boundary-aware loss. Through the ablation studies, the effectiveness of our designed modules is verified. We also evaluate the performance of our network by comparing it with other semantic BEV map generation methods. The results demonstrate the superiority of our network, especially for the performance in segmenting the minority classes. Although we achieve satisfying results, we find that the improvement in the prediction of minority classes may degrade the performance of the majority ones. In the future, we aim to deal with this problem by exploring the potential of the diffusion model to effectively capture the intricate distribution patterns from the dataset.

REFERENCES

[1] Y. Huang, J. Du, Z. Yang, Z. Zhou, L. Zhang, and H. Chen, "A survey on trajectory-prediction methods for autonomous driving," *IEEE Transactions on Intelligent Vehicles*, vol. 7, no. 3, pp. 652–674, 2022.

[2] J. Gu, C. Hu, T. Zhang, X. Chen, Y. Wang, Y. Wang, and H. Zhao, "Vip3d: End-to-end visual trajectory prediction via 3d agent queries," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5496–5506.

[3] Q. Meng, H. Guo, J. Li, Q. Dai, and J. Liu, "Vehicle trajectory prediction method driven by raw sensing data for intelligent vehicles," *IEEE Transactions on Intelligent Vehicles*, 2023.

[4] G. Aydemir, A. K. Akan, and F. Güney, "Adapt: Efficient multi-agent trajectory prediction with adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8295–8305.

[5] S. Teng, X. Hu, P. Deng, B. Li, Y. Li, Y. Ai, D. Yang, L. Li, Z. Xuanyuan, F. Zhu *et al.*, "Motion planning for autonomous driving: The state of the art and future perspectives," *IEEE Transactions on Intelligent Vehicles*, 2023.

[6] W. Ma, S. Huang, and Y. Sun, "Triplet-graph: Global metric localization based on semantic triplet graph for autonomous vehicles," *IEEE Robotics and Automation Letters*, vol. 9, no. 4, pp. 3155–3162, 2024.

[7] S. Teng, L. Chen, Y. Ai, Y. Zhou, Z. Xuanyuan, and X. Hu, "Hierarchical interpretable imitation learning for end-to-end autonomous driving," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 1, pp. 673–683, 2023.

[8] Y. Feng and Y. Sun, "Polarpoint-bev: Bird-eye-view perception in polar points for explainable end-to-end autonomous driving," *IEEE Transactions on Intelligent Vehicles*, 2024.

[9] X. Tang, K. Yang, H. Wang, J. Wu, Y. Qin, W. Yu, and D. Cao, "Prediction-uncertainty-aware decision-making for autonomous vehicles," *IEEE Transactions on Intelligent Vehicles*, vol. 7, no. 4, pp. 849–862, 2022.

[10] Y. Feng, W. Hua, and Y. Sun, "Nle-dm: Natural-language explanations for decision making of autonomous driving based on semantic scene understanding," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 9, pp. 9780–9791, 2023.

[11] X. He, H. Yang, Z. Hu, and C. Lv, "Robust lane change decision making for autonomous vehicles: An observation adversarial reinforcement learning approach," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 1, pp. 184–193, 2022.

[12] K. Mani, S. Daga, S. Garg, S. S. Narasimhan, M. Krishna, and K. M. Jatavallabhula, "Monolayout: Amodal scene layout from a single image," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 1689–1697.

[13] M. Saini and S. Susan, "Deep transfer with minority data augmentation for imbalanced breast cancer dataset," *Applied Soft Computing*, vol. 97, p. 106759, 2020.

[14] Y. Shi, T. ValizadehAslani, J. Wang, P. Ren, Y. Zhang, M. Hu, L. Zhao, and H. Liang, "Improving imbalanced learning by pre-finetuning with data augmentation," in *Fourth International Workshop on Learning with Imbalanced Domains: Theory and Applications*. PMLR, 2022, pp. 68–82.

[15] M. Temraz and M. T. Keane, "Solving the class imbalance problem using a counterfactual method for data augmentation," *Machine Learning with Applications*, vol. 9, p. 100375, 2022.

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[17] J. Li, Q. Zhu, Q. Wu, and Z. Fan, "A novel oversampling technique for class-imbalanced learning based on smote and natural neighbors," *Information Sciences*, vol. 565, pp. 438–455, 2021.

[18] Y. Xie, M. Qiu, H. Zhang, L. Peng, and Z. Chen, "Gaussian distribution based oversampling for imbalanced data classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 2, pp. 667–679, 2020.

[19] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," *Advances in neural information processing systems*, vol. 32, 2019.

[20] S. Rota Bulo, G. Neuhold, and P. Kontschieder, "Loss max-pooling for semantic image segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2126–2135.

[21] Z. Feng, Y. Guo, and Y. Sun, "Segmentation of road negative obstacles based on dual semantic-feature complementary fusion for autonomous driving," *IEEE Transactions on Intelligent Vehicles*, 2024.

[22] Y. Sun, W. Zuo, and M. Liu, "Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2576–2583, 2019.

[23] Z. Feng, Y. Guo, and Y. Sun, "Cekd: Cross-modal edge-privileged knowledge distillation for semantic scene understanding using only thermal images," *IEEE Robotics and Automation Letters*, vol. 8, no. 4, pp. 2205–2212, 2023.

[24] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[25] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[26] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[27] J. Fu, J. Liu, Y. Wang, J. Zhou, C. Wang, and H. Lu, "Stacked deconvolutional network for semantic segmentation," *IEEE Transactions on Image Processing*, 2019.

[28] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[29] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *European conference on computer vision*. Springer, 2022, pp. 205–218.

[30] C. Lu, M. J. G. van de Molengraft, and G. Dubbelman, "Monocular semantic occupancy grid mapping with convolutional variational encoder–decoder networks," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 445–452, 2019.

[31] B. Pan, J. Sun, H. Y. T. Leung, A. Andonian, and B. Zhou, "Cross-view semantic segmentation for sensing surroundings," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4867–4873, 2020.

[32] S. Gao, Q. Wang, and Y. Sun, "S2g2: Semi-supervised semantic bird-eye-view grid-map generation using a monocular camera for autonomous driving," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 11 974–11 981, 2022.

[33] J. Philion and S. Fidler, "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, 2020, pp. 194–210.

[34] Y. Li, Z. Ge, G. Yu, J. Yang, Z. Wang, Y. Shi, J. Sun, and Z. Li, "Bevdepth: Acquisition of reliable depth for multi-view 3d object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, 2023, pp. 1477–1485.

[35] S. H. Ebenuwa, M. S. Sharif, M. Alazab, and A. Al-Nemrat, "Variance ranking attributes selection techniques for binary classification problem in imbalance data," *IEEE access*, vol. 7, pp. 24 649–24 666, 2019.

[36] S. S. Mullick, S. Datta, and S. Das, "Adaptive learning-based k-nearest neighbor classifiers with resilience to class imbalance," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 11, pp. 5713–5725, 2018.

[37] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[38] D. Misra, T. Nalamada, A. U. Arasanipalai, and Q. Hou, "Rotate to attend: Convolutional triplet attention module," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 3139–3148.

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

[40] W. Yang, Q. Li, W. Liu, Y. Yu, Y. Ma, S. He, and J. Pan, "Projecting your view attentively: Monocular road scene layout estimation via cross-view transformation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15 536–15 545.