Models and Data Enhancements for Robust Map Segmentation in Autonomous Driving

Xingliang Huang
University of Chinese Academy of Sciences
huangxingliang20@mails.ucas.ac.cn

Yu Tian
Tsinghua University
tianyu1810613@gmail.com

Abstract

This technical report summarizes the winning solution for the track 2 (map segmentation) of the RoboDrive Challenge, which is affiliated with the 41st IEEE Conference on Robotics and Automation (ICRA 2024). Our proposed solution builds upon BEVerse, a camera-based baseline for the map segmentation task. We further study novel designs and optimization tailored to the robust map segmentation task, including perspective-view map loss, data augmentations, model scaling up, and effective post-processing strategies. These designs and optimization result in a state-of-the-art mIoU score of 48.75% on the corrupted nuScenes test set, ranking the 1st place in the challenge track 2.

1. Introduction

Sensing the map of the driving scene through a multi-view camera on the vehicle is a cost-effective and efficient solution in autonomous driving technology [1–4]. Map segmentation perception is important for path planning of autonomous driving systems by providing information about the boundaries of the road [5–7]. Camera-based solutions may have very fragile prediction results in case of bad weather, sensor noise, and other corruptions [8–11]. Therefore, the ability to perform accurate prediction tasks despite various damaged sensor images can greatly increase the safety of autonomous driving systems.

The track 2 of the 2024 RoboDrive Challenge [12] requires participants to develop map segmentation algorithms that solely utilize corrupted camera input during inference. In addition, it is not allowed to use corruption augmentation during training, which targets testing the Out-of-Distribution (OoD) robustness of the developed models. The impact of this challenge is significant because it provides a common corruption benchmark for autonomous driving perception in real-world corruption scenarios.

In this competition, we emphasize three aspects of methods, including data augmentation, model scale, and temporal post-process. This comes mainly from several motivations. First, limited training data restricts the model's generalization performance in corrupted scenarios. Designing an image augmentation strategy that applies to as many scenarios as possible can mitigate overfitting while also improving the model's performance under corrupted images [8, 13]. However, existing image augmentations mainly interfere with pixels through some predefined pixel values, such as zeros in Cutout [14] or Gaussian noise. In this technical report, we adopt a simple and effective data enhancement method that randomly shuffles a certain percentage of pixels, which makes the image enhancement more relevant to the input image rather than artificially defined. Besides, we have also explored vision backbones with different sizes under corrupted scenes. Last, we ensemble the predicted results at adjacent frames since map elements do not move with the vehicle. Predictions closer to the camera generally have more reliable robustness. As will be described in this report, strong data augmentation and post-processing strategy became the key factors for our success in this challenge.

2. Approach

In this section, we will present our solution in detail with the following aspects covered. Section 2.1 will elaborate on our model modification. Section 2.2 will discuss the data augmentation design. Finally, Section 2.3 will outline our post-processing strategies.

2.1. Model modification

Since BEVerse [1] is a multi-task visual perception algorithmic framework that uses shared BEV features to decode three tasks (3D object detection, map segmentation and motion prediction). To mitigate the performance loss due to shared parameters in multitasking, we retain the detection header for only one task, map segmentation. At the same time, the grid-aware range of BEV is adjusted to the range

Technical Report of the 2024 RoboDrive Challenge.
Track 2: Robust Map Segmentation.

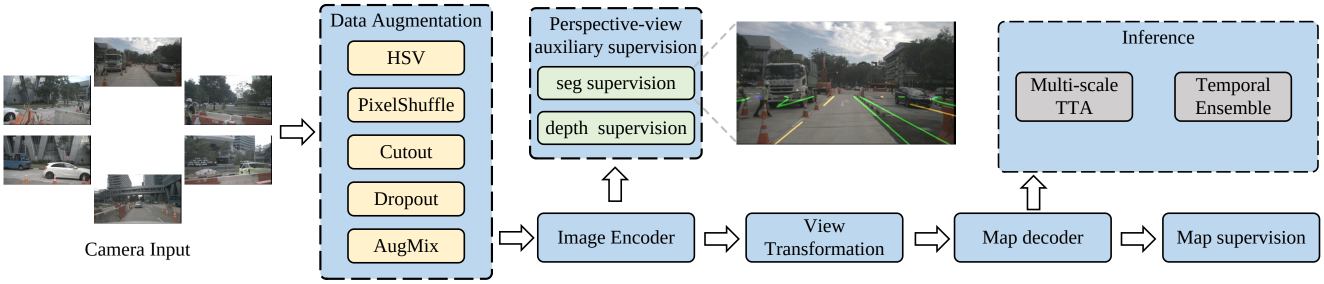

Figure 1. Illustration of the overall pipeline in our proposed robust map segmentation solution.

required by the map task. We adjusted the depth classification range of the LSS module [15] to 1 to 35 meters at 0.5m intervals.

BEVerse uses LSS as the perspective transformation module, and we adopt the module of depth supervision in BEVDepth into the baseline model. Camera parameters and image features are passed into a depth estimation network, and the Lidar point cloud is converted into depth information in the image to supervise the depth network. A cross-entropy loss function is used for the depth estimation loss $\mathcal{L}_{depth}$.

In order to guide the image encoder to learn rich map features, we project the map elements onto the camera plane and employ two convolutional layers as the segmentation head for the perspective view (PV) to predict the three types of map foregrounds. Dice loss and Cross-entropy loss are adopted as the auxiliary loss function:

$$\mathcal{L}_{PV} = \mathcal{L}_{bce}(\hat{M}_{PV}, M_{PV}) + \mathcal{L}_{dice}(\hat{M}_{PV}, M_{PV}). \quad (1)$$

$\hat{M}_{PV}$ is the predicted PV mask and M_{PV} is the ground truth. $\mathcal{L}_{bce}$ represents the binary cross-entropy loss and $\mathcal{L}_{dice}$ denotes dice loss [16].

Following BEVerse, the Cross-entropy loss function is employed as the loss for the map segmentation header. The overall loss of our solution is formulated as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{depth} + \lambda_2 \mathcal{L}_{PV} + \lambda_3 \mathcal{L}_{map} \quad (2)$$

where $\mathcal{L}_{map}$ is the map segmentation loss, and the weighted factors $\{\lambda_1, \lambda_2, \lambda_3\}$ are set to $\{1, 1, 10\}$ respectively.

2.2. Data augmentation

In this section, we present a simple yet effective data augmentation called PixelShuffle augmentation, which randomly shuffles a certain percentage of pixels of the original image. We first sample a ratio value from a specific range (0.1 $\sim$ 0.4). Then, using this ratio, a portion of the pixels in the image are randomly selected and spatially shuffled. The larger the proportion of random samples, the more pronounced the augmentation of the image. Figure 2 shows some augmented results at different ratios. As we can see in Figure 2, such an augmentation leads to better consistency between the processed image and the original one.

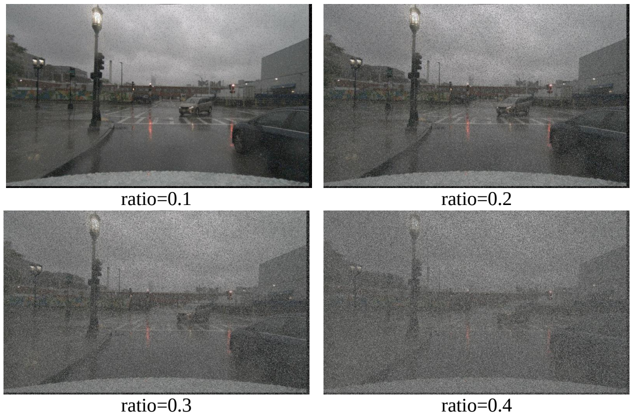

Figure 2. Examples of augmented images by PixelShuffle at different shuffle ratios.

In addition to PixelShuffle, several image-level data augmentation methods are employed, including HSV augmentation, Cutout [14], and Dropout. For each data augmentation, the final augmented image is obtained through AugMix, which blends the same image with different data augmentations. Following AugMix [17], a random selection of up to 3 augmentation methods is applied to the original image sequentially. Subsequently, the process was repeated three times to weight the three augmented images with coefficients sampled from the Dirichlet distribution. Finally, the weighted image is added to the original image using a parameter m controlled by the beta distribution:

$$I_{aug} = m \times I + (1 - m) \times I_{mix} \quad (3)$$

where I, I_{mix} and I_{aug} denote the original, weighted and augmented image, $m \in Beta(2, 6)$.

Multi-scale training is also adopted during training. We randomly resize the original images with a ratio uniformly sampled between 0.89 and 1.0. After that, the images are resized to a fixed size of 512 × 1408 pixels. Finally, all images are randomly flipped with a possibility of 0.5.

Table 1. Map segmentation performance of different settings on the RoboDrive track 2 test set.

Method	brightness	color quant	contrast	dark	defocus blur	elastic transform	fog	frost	gaussian noise	glass blur	impulse noise	iso noise	jpeg compression	motion blur	pixelate	shot noise	snow	zoom blur	mIoU
BEVerse	25.4	-	5.9	27.0	28.4	47.1	25.8	12.9	7.8	31.0	2.8	14.8	31.4	24.8	40.9	4.1	2.2	19.3	20.4
Single frame	33.4	-	9.1	35.4	32.5	44.6	28.0	6.3	16.8	30.8	12.9	27.6	29.0	17.2	39.3	11.5	9.5	19.3	23.5
+Depth	32.6	-	12.5	35.0	32.4	45.4	29.7	11.3	19.4	27.7	15.2	28.0	30.8	19.6	39.9	14.9	16.3	20.6	25.0
+HSV	50.3	-	26.0	41.3	36.1	45.9	31.0	8.6	26.7	29.3	19.2	35.0	34.3	19.6	38.2	17.6	26.9	15.9	29.1
+PixelShuffle	33.8	-	11.9	45.3	38.8	46.1	27.4	13.4	33.6	37.9	24.8	39.4	43.0	29.7	44.0	24.3	30.2	21.8	31.6
+AugMix	47.9	-	28.8	54.8	39.0	45.5	29.8	17.9	44.0	37.0	34.6	46.8	49.6	25.8	44.4	29.3	31.4	21.4	36.3
+PV seg	51.3	-	27.7	54.0	41.4	45.3	29.9	16.0	38.4	38.8	33.9	43.5	46.5	31.1	44.8	29.5	40.0	21.6	36.7
+Swin Large	48.5	-	45.3	58.5	48.8	45.9	31.0	19.2	47.5	38.2	37.6	50.2	51.0	35.2	44.3	31.8	45.8	20.7	40.9
+ Scale TTA	48.9	-	45.9	59.3	49.7	46.5	31.9	18.8	48.4	38.3	38.7	50.8	49.6	35.5	44.8	31.8	47.1	21.2	41.3
+ Temporal	54.6	-	54.6	71.1	64.8	52.1	28.6	23.1	58.5	51.2	46.1	64.2	54.9	44.7	55.2	37.2	54.5	21.8	48.8

2.3. Post-processing

We believe that post-processing techniques suitable for map segmentation can greatly enhance the results of map segmentation. We perform post-processing in spatial and temporal dimensions respectively.

2.3.1 Test-time augmentation

In the spatial dimension, multi-scale test is adopted during inference. The images are scaled several times at 0.82, 0.9, and 0.99 respectively. Then, we average the obtained map segmentation scores to get the final map segmentation result.

2.3.2 Temporal ensemble

The ensemble of temporal information can significantly enhance the robustness of the model. While BEVerse [1] is capable of fusing historical BEV features to facilitate map segmentation for the current frame, our findings indicate that this approach does not lead to an improvement in the model's robustness. Even the robustness of the model trained and evaluated on a single frame is higher than that observed when multiple frames are used. Consequently, in our solution, we train and test the model on a single frame only. The ensemble of temporal information is performed offline. Empirically, the accuracy of results closer to the vehicle tends to be higher. Therefore, we transform the results from both previous and future frames to the current frame. For static map grids, we leverage the predicted grids that are close to the ego car in multiple frames to replace the grids co-located in the current frame.

3. Experiments

In this section, we will present our experiments in detail with the following aspects. Section 3.1 provides details on the use of training and evaluation datasets. Section 3.2 elaborates on our experimental setups. Section 3.3 outlines our implementation details. Finally, Section 3.4 will discuss the comparative study and ablation study of our experiments.

3.1. Datasets

This work follows the protocol in the 2024 RoboDrive Challenge [12] when preparing the training and test data. Specifically, the model is trained on the official *train* split of the nuScenes dataset [18] and tested on the held-out competition evaluation sets. The evaluation data was created following RoboDepth [13, 19, 20], RoboBEV [8, 21, 22], and Robo3D [9, 23]. The corruption types are mainly from three sources, namely camera corruptions, camera failures, and LiDAR failures. For more details, please refer to the corresponding GitHub repositories.

3.2. Experimental Setups

The model is implemented based on the MMDetection3D codebase [24]. We use a batch size of 8 on 8 NVIDIA 4090 GPUs, AdamW optimizer with a peak learning rate of 2.5×10^{-4} and a weigh decay [25] of 0.01. We train our models around 20 epochs for map segmentation tasks. Cyclic learning rate schedule [26] is adopted during training. For the temporal ensemble window, we use 20 historical frames and 20 future frames. We only train the single-frame version of BEVerse [1], which is much more efficient and robust.

3.3. Implementation Details

The image scale is 512×1408 pixels. The image features from the backbone are downsampled with a stride of 16. We use commonly used data augmentation strategies, including flip and rotation on image space. The depth net predicts 68 discrete depth categories covering the depth from 1 m to 35 m. BEV features are transformed according to a grid size of 200×400 and a resolution of 0.15 meters. The vertical range in the LSS module is restricted from -1.5 m to 1.5 m.

3.4. Comparative Study

In our exploration, we first verify the effects of different models at a smaller scale and with fewer training epochs. In this setting, the image backbone is Swin-small [27] pre-trained on ImageNet [28], and the input scale is 512×1408 pixels. We list the milestones of our exploration in Table 1. BEVerse is our vanilla baseline. This baseline is trained with only 10 epochs on the nuScenes training set. We train a single-frame version of the baseline and find that the single-frame version shows better robustness with an improvement of 3.12 % mIoU. After incorporating depth supervision following BEVDepth, the robustness is further improved. Then we add HSV enhancement and PixelShuffle enhancement separately, and we get a huge improvement in robustness compared to that without augmentations. We further adopt AugMix to fuse all the augmentation operations, including HSV, PixelShuffle, Cutout, and Dropout. On top of that, we leverage perspective-view segmentation for the backbone features with an auxiliary loss. After verifying these effects of different models, we switch to a larger image backbone Swin-Large, and the training schedule is extended to 20 epochs. Finally, we adopt the test-time augmentation at different scales and ensemble the results from different timesteps.

Figure 3 and Figure 4 report some challenging examples from the RoboDrive Challenge. Even for severely corrupted images, our approach could predict accurate and consistent map segmentations, which demonstrates the strong robustness of our solution.

4. Conclusion

In this report, we describe our winning solution for the RoboDrive Challenge map segmentation in conjunction with ICRA 2024. Our solution demonstrates excellent map segmentation results. It also shows the effectiveness of data augmentations and temporal ensemble in the robustness of map segmentation.

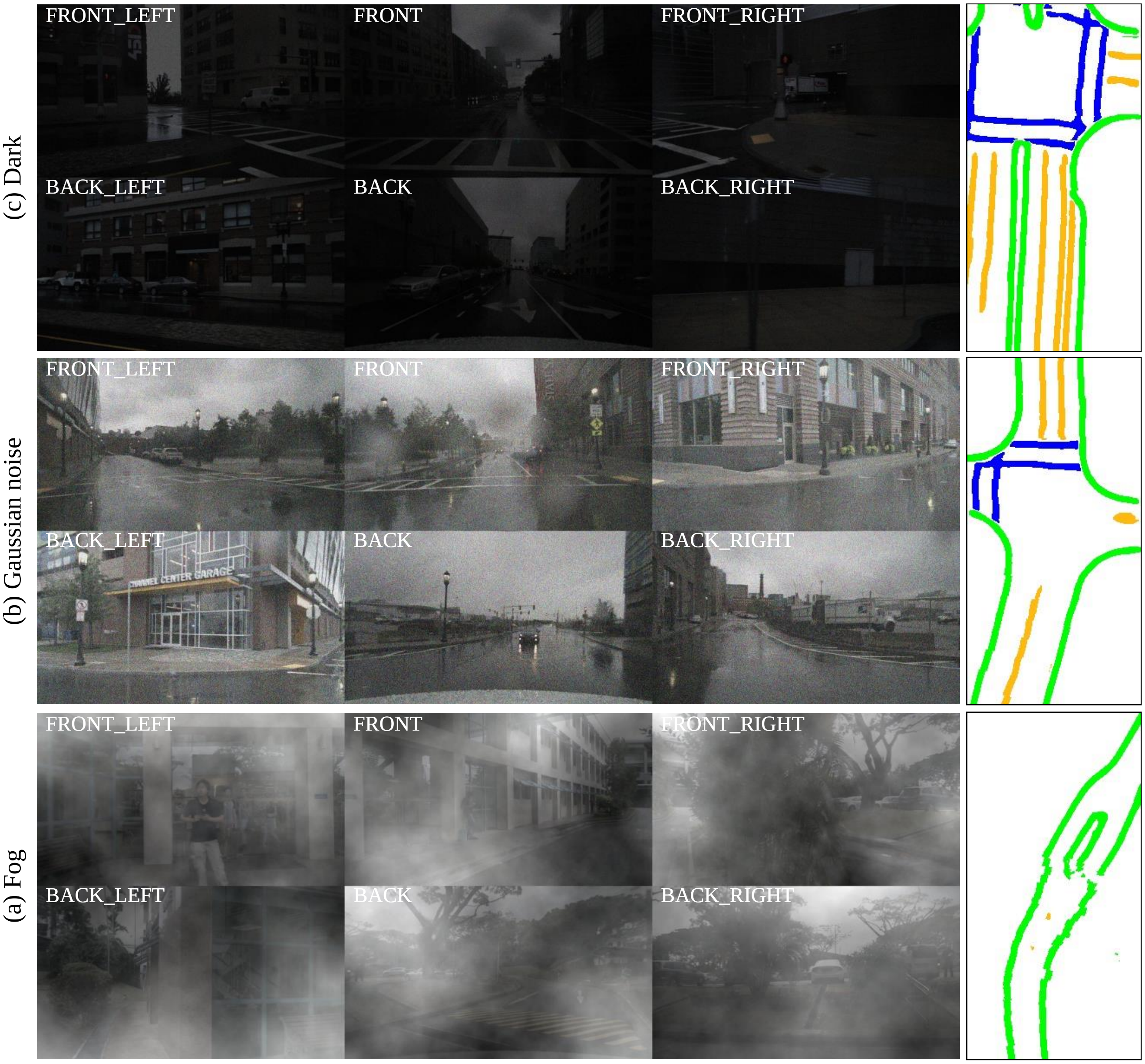

Figure 3. Qualitative results of our solution in the challenge under different corruptions (fog, gaussian noise, and dark).

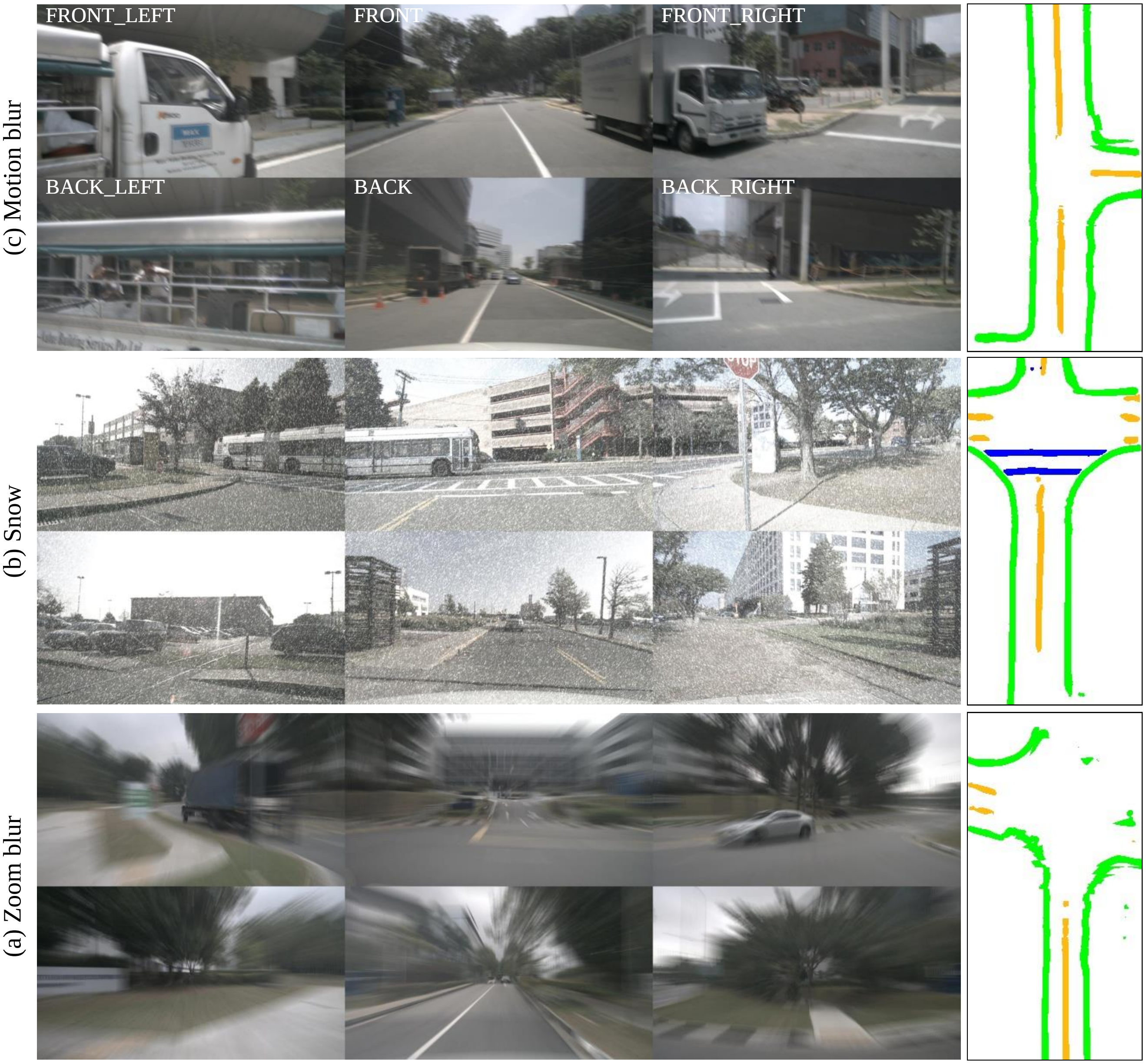

Figure 4. Qualitative results of our solution in the challenge under different corruptions (zoom blur, snow, and motion blur).

References

[1] Yunpeng Zhang, Zheng Zhu, Wenzhao Zheng, Junjie Huang, Guan Huang, Jie Zhou, and Jiwen Lu. Beverse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving. *arXiv preprint arXiv:2205.09743*, 2022.

[2] Lang Peng, Zhirong Chen, Zhangjie Fu, Pengpeng Liang, and Erkang Cheng. Bevsegformer: Bird's eye view semantic segmentation from arbitrary camera rigs. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5935–5943, 2023.

[3] Lingdong Kong, Jiawei Ren, Liang Pan, and Ziwei Liu. Lasermix for semi-supervised lidar semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21705–21715, 2023.

[4] Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, and Wenping Wang. Clip2scene: Towards label-efficient 3d scene understanding by clip. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7020–7030, 2023.

[5] Qi Li, Yue Wang, Yilun Wang, and Hang Zhao. Hdmapnet: An online hd map construction and evaluation framework. In *International Conference on Robotics and Automation*, pages 4628–4634, 2022.

[6] Lingdong Kong, Youquan Liu, Runnan Chen, Yuexin Ma, Xinge Zhu, Yikang Li, Yuenan Hou, Yu Qiao, and Ziwei Liu. Rethinking range view representation for lidar segmentation. In *IEEE/CVF International Conference on Computer Vision*, pages 228–240, 2023.

[7] Lingdong Kong, Niamul Quader, and Venice Erin Liong. Conda: Unsupervised domain adaptation for lidar segmentation via regularized domain concatenation. In *IEEE International Conference on Robotics and Automation*, pages 9338–9345, 2023.

[8] Shaoyuan Xie, Lingdong Kong, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Benchmarking and improving bird's eye view perception robustness in autonomous driving. *arXiv preprint arXiv:2405.17426*, 2024.

[9] Lingdong Kong, Youquan Liu, Xin Li, Runnan Chen, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Robo3d: Towards robust and reliable 3d perception against corruptions. In *IEEE/CVF International Conference on Computer Vision*, pages 19994–20006, 2023.

[10] Lingdong Kong, Xiang Xu, Jiawei Ren, Wenwei Zhang, Liang Pan, Kai Chen, Wei Tsang Ooi, and Ziwei Liu. Multimodal data-efficient 3d scene understanding for autonomous driving. *arXiv preprint arXiv:2405.05258*, 2024.

[11] Youquan Liu, Lingdong Kong, Jun Cen, Runnan Chen, Wenwei Zhang, Liang Pan, Kai Chen, and Ziwei Liu. Segment any point cloud sequences by distilling vision foundation models. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

[12] Lingdong Kong, Shaoyuan Xie, Hanjiang Hu, Yaru Niu, Wei Tsang Ooi, Benoit R. Cottereau, Lai Xing Ng, Yuexin Ma, Wenwei Zhang, Liang Pan, Kai Chen, Ziwei Liu, Weichao Qiu, Wei Zhang, Xu Cao, Hao Lu, Ying-Cong Chen, Caixin Kang, Xinning Zhou, Chengyang Ying, Wentao Shang, Xingxing Wei, Yinpeng Dong, Bo Yang, Shengyin Jiang, Zeliang Ma, Dengyi Ji, Haiwen Li, Xingliang Huang, Yu Tian, Genghua Kou, Fan Jia, Yingfei Liu, Tiancai Wang, Ying Li, Xiaoshuai Hao, Yifan Yang, Hui Zhang, Mengchuan Wei, Yi Zhou, Haimei Zhao, Jing Zhang, Jinke Li, Xiao He, Xiaoqiang Cheng, Bingyang Zhang, Lirong Zhao, Dianlei Ding, Fangsheng Liu, Yixiang Yan, Hongming Wang, Nanfei Ye, Lun Luo, Yubo Tian, Yiwei Zuo, Zhe Cao, Yi Ren, Yunfan Li, Wenjie Liu, Xun Wu, Yifan Mao, Ming Li, Jian Liu, Jiayang Liu, Zihan Qin, Cunxi Chu, Jialei Xu, Wenbo Zhao, Junjun Jiang, Xianming Liu, Ziyan Wang, Chiwei Li, Shilong Li, Chendong Yuan, Songyue Yang, Wentao Liu, Peng Chen, Bin Zhou, Yubo Wang, Chi Zhang, Jianhang Sun, Hai Chen, Xiao Yang, Lizhong Wang, Dongyi Fu, Yongchun Lin, Huitong Yang, Haoang Li, Yadan Luo, Xianjing Cheng, and Yong Xu. The robodrive challenge: Drive anytime anywhere in any condition. *arXiv preprint arXiv:2405.08816*, 2024.

[13] Lingdong Kong, Shaoyuan Xie, Hanjiang Hu, Lai Xing Ng, Benoit Cottereau, and Wei Tsang Ooi. Robodepth: Robust out-of-distribution depth estimation under corruptions. *Advances in Neural Information Processing Systems*, 36, 2024.

[14] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[15] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European Conference on Computer Vision*, pages 194–210. Springer, 2020.

[16] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *International Conference on 3D Vision*, pages 565–571, 2016.

[17] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019.

[18] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020.

[19] Lingdong Kong, Shaoyuan Xie, Hanjiang Hu, Benoit Cottereau, Lai Xing Ng, and Wei Tsang Ooi. The robodepth benchmark for robust out-of-distribution depth estimation under corruptions. `https://github.com/ldkong1205/RoboDepth`, 2023.

[20] Lingdong Kong, Yaru Niu, Shaoyuan Xie, Hanjiang Hu, Lai Xing Ng, Benoit Cottereau, Ding Zhao, Liangjun Zhang, Hesheng Wang, Wei Tsang Ooi, Ruijie Zhu, Ziyang Song, Li Liu, Tianzhu Zhang, Jun Yu, Mohan Jing, Pengwei Li, Xiaohua Qi, Cheng Jin, Yingfeng Chen, Jie Hou, Jie Zhang, Zhen Kan, Qiang Lin, Liang Peng, Minglei Li, Di Xu, Changpeng Yang, Yuanqi Yao, Gang Wu, Jian Kuai, Xianming Liu, Junjun Jiang, Jiamian Huang, Baojun Li, Jiale Chen, Shuang Zhang, Sun Ao, Zhenyu Li, Runze Chen, Haiyong Luo, Fang Zhao, and Jingze Yu. The robodepth challenge: Methods and advancements towards robust depth estimation. *arXiv preprint arXiv:2307.15061*, 2023.

[21] Shaoyuan Xie, Lingdong Kong, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Robobev: Towards robust bird's eye view perception under corruptions. *arXiv preprint arXiv:2304.06719*, 2023.

[22] Shaoyuan Xie, Lingdong Kong, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. The robobev benchmark for robust bird's eye view detection under common corruption and domain shift. https://github.com/Daniel-xsy/RoboBEV, 2023.

[23] Lingdong Kong, Youquan Liu, Xin Li, Runnan Chen, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. The robo3d benchmark for robust and reliable 3d perception. https://github.com/ldkong1205/Robo3D, 2023.

[24] MMDetection3D Contributors. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. https://github.com/open-mmlab/mmdetection3d, 2020.

[25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[26] Leslie N Smith. Cyclical learning rates for training neural networks. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 464–472, 2017.

[27] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 2012.