MultiViewRobust: Scaling Up Pretrained Models for Robust Map Segmentation

Genghua Kou
Beijing Institute of Technology
Beijing, China
koughua@bit.edu.cn

Fan Jia
Megvii Technology
Beijing, China
jiafan@megvii.com

Yingfei Liu
Megvii Technology
Beijing, China
liuyingfei@megvii.com

Tiancai Wang
Megvii Technology
Beijing, China
wangtiancai@megvii.com

Ying Li
Beijing Institute of Technology
Beijing, China
ying.li@bit.edu.cn

Abstract

Although existing map segmentation methods have achieved high performance, they struggle to handle challenges introduced by real-world corruption. The CrazyFriday team focuses on the scaling up framework to improve the generalized ability of robustness, and developed an enhanced MultiViewRobust framework to address the challenge of robust map segmentation, leveraging multi-view architecture and advanced temporal information integration. We achieved second place in the track-2 of the RoboDrive Challenge.

1. Introduction

Map segmentation is essential for autonomous driving tasks such as HD map construction, which is crucial for driving safety [1–10]. However, various challenging scenes can affect the accuracy of map segmentation. These scenes are rare in *clean* datasets but often occur in the real world, leading to the performance dropping of some high-performance approaches [11]. These approaches tend to over-fit certain datasets, which may lead to poor performance of robustness [12]. Fortunately, the RoboDrive competition [13] provides datasets and toolkits for training and testing the robustness of frameworks.

To address these problems, we initially conducted experiments on various recent high-performance models [14–17] to compare their abilities. Temporal and multi-view fusion strategy is widely implemented in these models to achieve robust map segmentation against corruptions [18–20]. Therefore, we selected BEVerse [14] as the baseline.

Next, we attempted to address factors that may affect performance or robustness under corrupted images. The module of the branch for a specific task may contribute little to robustness. These specific task branches mainly focus on the specific task of refinement, receiving features from the backbone. The backbone processes the image primarily to reconstruct the features under corruption. Therefore, the robustness of the framework may be attributed to the well-informed backbone [21–24] for the generalized ability of robust image feature extraction.

Finally, we proposed an enhanced framework named MultiViewRobust, which uses enhanced backbone integration, temporal and multi-view fusion, advanced post-processing techniques, and some training strategies for map segmentation under corrupted images. The MultiViewRobust achieved second place in the track-2 of the RoboDrive Challenge, demonstrating the effectiveness of our framework.

2. Approach

Given a surrounding image I, our framework first extracts features from the image. To enhance the robustness of feature extraction, we utilize the large image backbone of Swin-L [24] or EVA-02 [22]. Following BEVerse [14], the multi-level features of the backbone are used to enable efficient fusion.

The inclusion of temporal information can help alleviate corruption []. Therefore, we employ image-to-BEV transformation [25] to convert perspective image features into a dense point cloud with various depths and camera intrinsic and extrinsic. For each timestamp, the view transformer utilizes multi-view features to cover the entire surroundings. Additionally, pillar pooling [26] is applied to these point clouds to create the BEV feature representation. These ef-

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

fectively handle temporal discrepancies and leverage spatial context, enhancing map segmentation accuracy.

After the view transformation, in line with FIERY [27], we first align the BEV features from past timestamps to the present reference frame using known ego motions. The aligned 4D features are then processed with a spatio-temporal BEV encoder to further extract spatial and temporal information. These refine the map segmentation outputs, ensuring high fidelity in the representation of dynamic and complex urban environments.

Finally, map decoders are employed for semantic map construction, utilizing a simple MLP.

3. Experiments

3.1. Datasets

This work follows the protocol in the 2024 RoboDrive Challenge [13] when preparing the training and test data. Specifically, the model is trained on the official *train* split of the nuScenes dataset [1] and tested on the held-out competition evaluation sets. The evaluation data was created following RoboDepth [28–30], RoboBEV [11, 31, 32], and Robo3D [12, 33]. The corruption types are mainly from three sources, namely camera corruptions, camera failures, and LiDAR failures.

3.2. Implementation Details

Apart from the basic setup of BEVerse [14], we select EVA-02-L [22] as the backbone pretrained on ImageNet-21K [34] for training our framework. The pretrained weights can be downloaded from this url. The input image size is 800×1600 pixels. We use AdamW [35] with a learning rate of 1e-5 and weight decay of 1e-2. The batch size is set to 1 during training. The entire model is trained for approximately 12 epochs on a server with 8 NVIDIA A100 GPUs.

3.3. Comparative Study

As shown in Tab. 1, we conducted a comparative analysis of different backbones, including Swin-S [23], Swin-L [23], and EVA-02-L [22]. Scaling up vision ability demonstrates the attribution to improved performance by increasing the number of parameters and data. Larger parameters and a pretrained dataset for the backbone may imply improving robustness of feature extraction under various corruptions.

3.4. Ablation Study

As shown in Fig. 1, we also investigated the effect of training epochs. Extending it leads to a drop in mIOU. We hypothesize that the model may overfit the dataset due to the optimizer focusing on a specific refinement strategy. This refinement may affect the robustness under varied and challenging conditions, while improving performance on specific datasets.

Backbone	Dataset	Parameters	mIOU
Swin-S [23]	ImageNet-1K [34]	50M	15.67
Swin-L [23]	ImageNet-21K [34]	197M	17.51
EVA-02-L [22]	ImageNet-21K [34]	304M	34.54

Table 1. Results of evaluating different pretrained backbones on the test servers provided by RoboDrive.

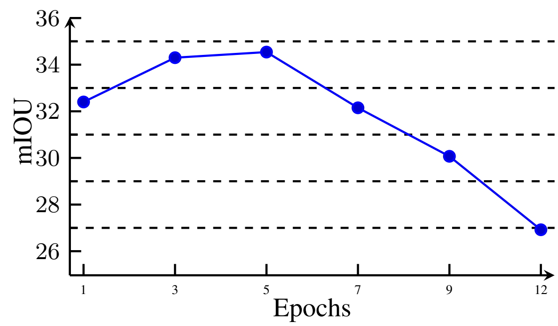

Figure 1. Illustration of the mIOU corresponding to the training epochs.

4. Conclusion

In this work, based on the high-performance framework we proposed an enhanced framework named MultiViewRobust, which uses enhanced backbone integration, temporal and multi-view fusion, and advanced post-processing techniques for map segmentation. the CrazyFriday team focuses on scaling up and the training strategy to improve the robustness under various corruption scenarios. We achieved second-place performance in the second track of the RoboDrive Challenge.

References

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

[2] Bencheng Liao, Shaoyu Chen, Xinggang Wang, Tianheng Cheng, Qian Zhang, Wenyu Liu, and Chang Huang. Maptr: Structured modeling and learning for online vectorized hd map construction. *ICLR*, 2022.

[3] 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.

[4] 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.

[5] 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.

[6] Sheng Yang, Xiaoling Zhu, Xing Nian, Lu Feng, Xiaozhi Qu, and Teng Ma. A robust pose graph approach for city scale lidar mapping. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1175–1182, 2018.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *ECCV*, pages 531–548, 2022.

[16] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Aqi Gao, Tiancai Wang, and Xiangyu Zhang. Petrv2: A unified framework for 3d perception from multi-camera images. In *ICCV*, pages 3239–3249, 2023.

[17] Jiacheng Chen, Yuefan Wu, Jiaqi Tan, Hang Ma, and Yasutaka Furukawa. Maptracker: Tracking with strided memory fusion for consistent vector hd mapping. *arXiv preprint arXiv:2403.15951*, 2024.

[18] Junjie Yan, Yingfei Liu, Jianjian Sun, Fan Jia, Shuailin Li, Tiancai Wang, and Xiangyu Zhang. Cross modal transformer: Towards fast and robust 3d object detection. In *ICCV*, pages 18268–18278, 2023.

[19] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *ICCV*, pages 3598–3608, 2023.

[20] Xiaohui Jiang, Shuailin Li, Yingfei Liu, Shihao Wang, Fan Jia, Tiancai Wang, Lijin Han, and Xiangyu Zhang. Far3d: Expanding the horizon for surround-view 3d object detection. In *AAAI*, volume 38, pages 2561–2569, 2024.

[21] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *CVPR*, pages 19358–19369, 2023.

[22] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *arXiv preprint arXiv:2303.11331*, 2023.

[23] 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 *ICCV*, 2021.

[24] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In *CVPR*, 2022.

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

[26] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *CVPR*, pages 12697–12705, 2019.

[27] Anthony Hu, Zak Murez, Nikhil Mohan, Sofía Dudas, Jeffrey Hawke, Vijay Badrinarayanan, Roberto Cipolla, and Alex Kendall. Fiery: Future instance prediction in bird's-eye view from surround monocular cameras. In *ICCV*, pages 15273–15282, 2021.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2017.