FORTRESS: Feature Optimization and Robustness Techniques for 3D Object Detection Systems

Caixin Kang[1], Xinning Zhou[2], Chengyang Ying[2], Wentao Shang[4],
Shiji Zhao[1], Xingxing Wei[1,3]*, Yinpeng Dong[2]*, Hang Su[2]*
[1] Institute of Artificial Intelligence, Beihang University
[2] Dept. of Comp. Sci. & Tech., Institute for AI, BNRist Center, THBI Lab,
Tsinghua-Bosch Joint ML Center, Tsinghua University
[3] Hangzhou Innovation Institute, Beihang University [4] Hefei University of Technology
{caixinkang,xxwei}@buaa.edu.cn, {dongyinpeng,suhangss}@tsinghua.edu.cn

Abstract

*The **FORTRESS** (Feature Optimization and Robustness Techniques for 3D Detection Systems) method introduces a novel approach to enhancing the robustness of 3D object detection in autonomous driving. Building on the RayDN architecture, FORTRESS incorporates a modified EVA ViT-Large backbone, pre-trained on ImageNet, to achieve deep and resilient feature extraction. The method is further enhanced with a strategic combination of Augmix and DeepAug data augmentation techniques, carefully crafted to address diverse environmental changes and maintain robustness against real-world data distribution shifts. The training process is systematically structured, progressing from clean datasets to increasingly complex scenarios, each phase contributing to the development of a more robust detection system. By adopting feature optimization and robustness techniques, FORTRESS not only refines the detection capabilities but also ensures the model's adaptability to varied and unforeseen environmental conditions. Preliminary results have demonstrated the method's potential as an effective solution for robust BEV detection challenges in autonomous driving. Additionally, FORTRESS was validated in the ICRA 2024 RoboDrive Challenge, where it achieved second place in Track 1: Robust BEV Detection.*

1. Introduction

The rise of autonomous driving technologies has intensified the focus on developing robust detection systems that can accurately interpret and navigate complex environments. Bird's Eye View (BEV) detection is particularly critical, offering a comprehensive perspective essential for the safe operation of autonomous vehicles. The RoboDrive Challenge, specifically Track 1: Robust BEV Detection, provides a platform to address this challenge using advanced computer vision techniques. Our solution, FORTRESS (Feature Optimization and Robustness Techniques for 3D Detection Systems), aims to significantly improve the accuracy and robustness of BEV detection.

BEV detection is vital for understanding a vehicle's surroundings from a top-down perspective, integrating sensor data to create a comprehensive view of road conditions, obstacles, and navigational cues. However, the dynamic nature of driving environments and the limitations of current detection technologies present significant challenges, such as varying environmental conditions, occlusions, and the need for precise object detection and depth estimation.

FORTRESS addresses these challenges by incorporating a novel pipeline for camera-only 3D object detection, a sophisticated feature extraction backbone, and data augmentation techniques. This approach ensures performance and adaptability in real-world scenarios, setting a new benchmark for BEV detection systems in autonomous vehicles.

2. Related Work

2.1. 3D Object Detection

The field of 3D object detection has advanced significantly with the rise of deep learning and computer vision. Initial approaches relied on geometric properties and stereoscopic vision for depth estimation [18]. However, the introduction of deep convolutional neural networks shifted focus towards data-driven methods, enhancing model accuracy through large-scale training. Notable milestones include LiDAR-based models like PointNet and PointRCNN, which have set benchmarks for 3D detection accuracy [14, 15]. More recently, cost-effective methods that infer 3D information

*Corresponding authors.

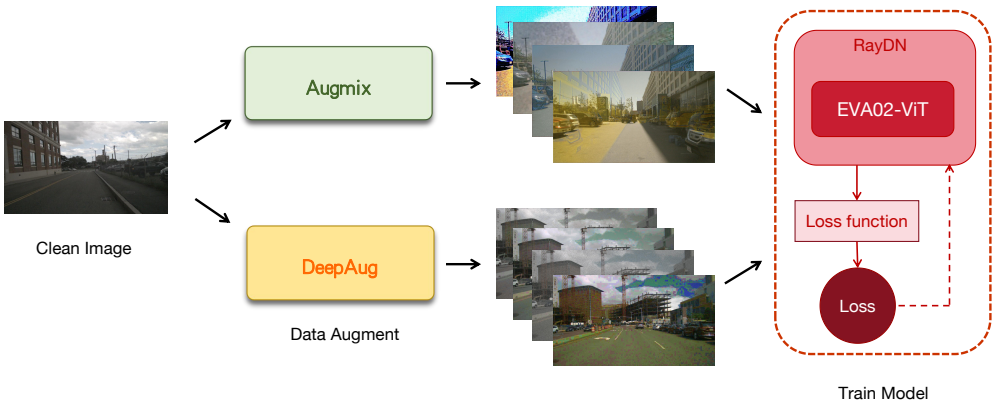

Figure 1. Pipeline of FORTRESS.

from 2D images, such as MonoDIS and Pseudo-LiDAR, have gained traction due to their seamless integration with camera systems [16, 17]. Advanced techniques for multi-view camera 3D detection in BEV space, like BEVFormer, BEVDepth, and Sparse4D, have achieved performance levels comparable to LiDAR-based approaches, marking substantial progress [9–11].

2.2. Robustness of Visual Systems

Research on adversarial robustness has focused on defending models against malicious inputs designed to cause errors, with foundational work by Madry et al. and Goodfellow et al. demonstrating the effectiveness of adversarial training [5, 7, 13]. Meanwhile, natural robustness, crucial for real-world deployment, addresses a model's reliability across diverse environmental conditions and sensor noise. Enhancements in natural robustness often involve data augmentation and robust training that simulate real-world disturbances. Benchmarks like ImageNet-C have been instrumental in testing models against common visual corruptions, driving the development of more resilient architectures [2, 6, 8].

3. Methodology

Our method for the RoboDrive Challenge, titled FORTRESS (Feature Optimization and Robustness Techniques for 3D Detection Systems), leverages advanced computational techniques to enhance the robustness and accuracy of Bird's Eye View (BEV) detection. By integrating sophisticated data processing, augmentation, and adaptive training strategies, FORTRESS effectively addresses the challenges of 3D object detection in dynamic driving environments. The pipeline of FORTRESS is illustrated in Fig. 1.

3.1. Pipeline

FORTRESS follows a novel pipeline based on RayDN [12] for camera-only 3D object detection, with specific enhancements tailored for multi-view scenarios. This method mitigates common issues like redundant and incorrect detections, which often arise from challenges in depth estimation from 2D images. By implementing depth-aware hard negative sampling along camera rays, the method generates hard negative examples that are visually similar to true positives, forcing the model to improve its depth-related feature discrimination. This plug-and-play module integrates seamlessly with any DETR-style multi-view 3D detector, offering a significant boost in detection accuracy without increasing computational overhead or affecting inference speeds, as demonstrated by its superior performance on the NuScenes dataset [1].

3.2. Backbone

FORTRESS utilizes the EVA ViT-Large [3], a next-generation Transformer-based model pre-trained on ImageNet. The EVA-02 [4] variant employs a plain Transformer architecture extensively trained to reconstruct robust, language-aligned vision features via masked image modeling. This backbone excels at extracting high-quality features crucial for precise object detection under variable environmental conditions, all while maintaining efficiency, making it ideal for autonomous driving applications.

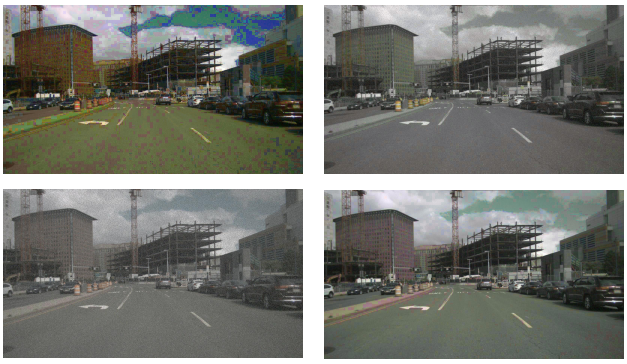

Figure 2. Visualization of Augmix-enhanced Data.

Figure 3. Visualization of DeepAug-enhanced Data.

3.3. Data Augmentation

To ensure reliability across diverse conditions, FORTRESS incorporates efficient data augmentation strategies: Augmix and DeepAug. Visualization of enhanced data can be seen in Fig. 2 and Fig. 3.

Augmix enhances model robustness by applying a sequence of image processing techniques—such as pixel shuffle, random hue, and random saturation—that preserve the semantic integrity of images while introducing diverse, realistic variations. This approach significantly improves the model's ability to estimate uncertainty and resist data corruptions that were not present during training, effectively bridging the gap between controlled training conditions and real-world scenarios.

The method constructs augmented images by applying a series of operations to the original image x. Let $\mathcal{O}_1, \mathcal{O}_2, \ldots, \mathcal{O}_n$ represent these operations. An augmentation chain can be expressed as:

$$x' = \mathcal{O}_n(\mathcal{O}_2(\mathcal{O}_1(x)) \ldots)$$

The outputs from multiple chains $x'_1, x'_2, \ldots, x'_k$ are then combined using element-wise convex combinations, with mixing weights $\mathbf{w}$ sampled from a Dirichlet distribution $\text{Dir}(\alpha, \ldots, \alpha)$. The mixed image $\tilde{x}$ is represented as:

$$\tilde{x} = w_1 \cdot x'_1 + w_2 \cdot x'_2 + \cdots + w_k \cdot x'_k$$

Finally, $\tilde{x}$ is blended with the original image x using another convex combination, where the mixing coefficient β is sampled from a Beta distribution $\text{Beta}(\alpha, \alpha)$. The resulting image y is defined as:

$$y = \beta \cdot x + (1 - \beta) \cdot \tilde{x}$$

This comprehensive process, incorporating randomness in operation selection, severity, and mixing, enhances the model's robustness and generalization, preparing it to handle a wide range of unseen data variations and corruptions.

DeepAug introduces a more innovative approach to data augmentation by focusing on the internal representations within deep neural networks rather than applying transformations directly to raw images. Clean images are processed through image-to-image networks, such as CAE and EDSR, where random perturbations are introduced at various network layers. This process generates images that, while preserving semantic integrity, exhibit significant visual differences from their original counterparts. Perturbations include operations like zeroing, negating, and convolving, which create diverse visual variations. This approach effectively trains the model to recognize and adapt to a wider spectrum of visual inputs.

The combination of Augmix and DeepAug equips FORTRESS to handle a broad range of environmental changes and ensures robustness against shifts in data distributions encountered during real-world deployment.

3.4. Training Strategy

The training strategy for FORTRESS is carefully structured to expose the model to a diverse array of scenarios. It begins with training on clean, unaltered data from the nuScenes dataset to establish a baseline for accuracy and robustness. As training progresses, complexity is introduced incrementally: first by incorporating data enhanced with Augmix, and then by integrating data augmented with both Augmix and DeepAug. This phased approach not only layers the model's robustness but also ensures it develops the ability to generalize effectively across varied environmental and operational conditions, ultimately leading to a more resilient and reliable detection system.

Through these strategic implementations, FORTRESS establishes a new approach for robustness and accuracy in BEV detection, offering a comprehensive and adaptable solution to meet the evolving challenges of autonomous driving technology.

4. Experiments

4.1. Experimental Setups

Our proposed approach was implemented using the PyTorch framework, with the FORTRESS model trained on eight NVIDIA GeForce RTX 4090 GPUs. The training utilized only the images from the training split of the nuScenes dataset. The regimen began with 24 epochs on the clean nuScenes training data, followed by 16 epochs with Augmix-enhanced data, and concluded with 16 epochs on data augmented with both Augmix and DeepAug.

4.2. Implementation Details

Augmix. The initial implementation of Augmix shared substantial overlap with the corruptions used in the RoboDrive competition. Given the competition rules prohibiting the use of identical corruptions during training, we propose alternative augmentation techniques—specifically pixel shuf-

Table 1. NDS of corruption categories on the Robodrive Challenge Track1 phase2 test dataset. (Part 1)

Corruptions	Bright	Dark	Fog	Frost	Snow	Contrast	Defocus Blur	Glass Blur	Motion Blur	Zoom Blur
RayDN	0.354	0.528	0.334	0.256	0.616	0.336	0.493	0.451	0.380	0.119
FORTRESS (Augmix)	0.421	0.627	0.336	0.439	0.648	0.480	0.587	0.434	0.413	0.156
FORTRESS (Augmix+DeepAug)	0.431	0.627	0.375	0.466	0.609	0.492	0.584	0.465	0.447	0.188

Table 2. NDS of corruption categories on the Robodrive Challenge Track1 phase2 test dataset. (Part 2)

Corruptions	Elastic Transform	Color Quant	Gaussian Noise	Impluse Noise	Shot Noise	ISO Noise	Pixelate	JPEG	**Average**
RayDN	0.470	0.487	0.588	0.363	0.483	0.482	0.559	0.429	0.429
FORTRESS (Augmix)	0.434	0.661	0.691	0.468	0.532	0.511	0.566	0.382	0.488
FORTRESS (Augmix+DeepAug)	0.448	0.667	0.706	0.424	0.560	0.508	0.568	0.475	0.502

Table 3. Clean performance on the nuScenes dataset validation split.

Models	NDS	mAP	mATE	mASE	mAOE	mAVE	mAAE
BEVFormer	0.517	0.415	0.672	0.274	0.369	0.397	0.198
RayDN	0.624	0.541	0.518	0.252	0.274	0.230	0.195
FORTRESS (Augmix)	0.623	0.541	0.509	0.253	0.268	0.248	0.194
FORTRESS (Augmix+DeepAug)	0.619	0.536	0.506	0.256	0.294	0.248	0.187

fle, random hue, and random saturation. These methods were chosen to simulate data degradation while adhering to the rules, thus enhancing the detection model's generalization capabilities.

DeepAug. DeepAug involves augmenting data using CAE and EDSR models. This process is integrated during the image loading phase, either on-the-fly or using pre-generated augmented data to optimize computational efficiency. In practice, we pre-generate the augmented data, and during training, images are loaded based on a random probability rd (with a threshold t empirically set to 0.6). If rd exceeds t, EDSR-processed data is used; otherwise, CAE-processed data is employed. To maintain consistency in detection outcomes, certain operations typically included in DeepAug, such as horizontal and vertical flips, were excluded.

4.3. Comparative Study

The RoboDrive Track 1: Robust BEV Detection competition challenges algorithms to recover from 18 types of corruptions, each designed to simulate various environmental and sensor-induced damages. Tab. 1 and Tab. 2 present the results of our baseline RayDN model and the FORTRESS method across these 18 corruption types, evaluated using the NDS metric. After applying Augmix, we observed performance improvements on most corruption types, including notable gains of 0.182 and 0.173 NDS on Frost and Color Quant, respectively, resulting in an average NDS of 0.488. However, slight performance declines were noted on corruptions such as Glass Blur, Elastic Transform, and JPEG. With the subsequent incorporation of DeepAug, overall robustness further improved, raising the average NDS to 0.502. These results demonstrate that both Augmix and DeepAug contribute significantly to enhancing NDS across the dataset.

4.4. Results on the nuScenes Dataset

We also evaluated the performance of our methods on the clean data from the nuScenes validation split, as shown in Tab. 3. RayDN demonstrated a notable improvement, achieving an NDS increase of 0.1068 compared to BEVFormer, establishing a strong foundation for our approach. The FORTRESS method, after incorporating Augmix augmentation, maintained nearly all of its performance on clean data. Following an additional 16 epochs of training with DeepAug-enhanced data, the NDS on clean data slightly decreased to 0.619. However, at this stage, FORTRESS achieved the highest robustness NDS, indicating a trade-off between performance on clean and robust data. This result underscores that FORTRESS's augmentation strategies effectively enhance robustness without significantly compromising performance on uncorrupted datasets, preserving its detection capabilities.

5. Conclusion

In this study, we presented FORTRESS, an advanced approach aimed at enhancing the robustness and accuracy of BEV detection for autonomous vehicles. By integrating Ray Denoising with the EVA ViT-Large backbone and leveraging effective data augmentation techniques such as Augmix and DeepAug, FORTRESS significantly improves detection performance across diverse environmental conditions. Our results demonstrated substantial gains in handling various data corruptions in the RoboDrive Challenge, while maintaining strong performance on clean data from the nuScenes dataset. This work establishes a foundation for further research into reliable and efficient BEV detection systems, striving to achieve a balance between high performance and robustness in real-world scenarios.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 2

[2] Yinpeng Dong, Caixin Kang, Jinlai Zhang, Zijian Zhu, Yikai Wang, Xiao Yang, Hang Su, Xingxing Wei, and Jun Zhu. Benchmarking robustness of 3d object detection to common corruptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1022–1032, 2023. 2

[3] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369, 2023. 2

[4] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171, 2024. 2

[5] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 2

[6] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 2

[7] Caixin Kang, Yinpeng Dong, Zhengyi Wang, Shouwei Ruan, Hang Su, and Xingxing Wei. Diffender: Diffusion-based adversarial defense against patch attacks in the physical world. *arXiv preprint arXiv:2306.09124*, 2023. 2

[8] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19994–20006, 2023. 2

[9] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1477–1485, 2023. 2

[10] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, pages 1–18. Springer, 2022.

[11] Xuewu Lin, Tianwei Lin, Zixiang Pei, Lichao Huang, and Zhizhong Su. Sparse4d: Multi-view 3d object detection with sparse spatial-temporal fusion. *arXiv preprint arXiv:2211.10581*, 2022. 2

[12] Feng Liu, Tengteng Huang, Qianjing Zhang, Haotian Yao, Chi Zhang, Fang Wan, Qixiang Ye, and Yanzhao Zhou. Ray denoising: Depth-aware hard negative sampling for multi-view 3d object detection, 2024. 2

[13] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 2

[14] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 1

[15] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 770–779, 2019. 1

[16] Andrea Simonelli, Samuel Rota Bulo, Lorenzo Porzi, Manuel López-Antequera, and Peter Kontschieder. Disentangling monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1991–1999, 2019. 2

[17] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8445–8453, 2019. 2

[18] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4490–4499, 2018. 1