A Neural Network Model for Autonomous Vehicles Safety Check Using Camera-only/ Camera-Lidar Images

Dengyi Liu[1], Honggang Wang[1], and Hua Fang[2]
[1] Graduate Computer Science and Engineering Department,
Katz School of Health and Science, Yeshiva University NY 10016
[3]Dept. of Computer and Information Science, University of Massachusetts Dartmouth, MA 02747
Emails: dliu6@mail.yu.edu, honggang.wang@yu.edu, and hfang2@umassd.edu

***Abstract*—Camera and LiDAR sensors play a crucial role in vehicle perception systems, enabling accurate detection of obstacles and other vehicles in autonomous driving technologies. Despite their complementary advantages, ensuring the reliability of camera data under varying environmental conditions poses a significant challenge. This paper introduces a novel system for camera data consistency checking through two methods: (1) a camera-only consistency checking mechanism utilizing an effective point-to-point feature extraction neural network inspired by the D2-Net method, and (2) a camera-LiDAR fusion data consistency checking approach employing advanced sensor fusion techniques. By integrating a critical reliability check for camera data and fusing it with LiDAR information, our system leverages the strengths of both sensor types to enhance the robustness and efficiency of autonomous vehicle navigation. Experimental results demonstrate that the proposed methods significantly improve the safety and performance of autonomous driving by ensuring reliable sensor data fusion, enabling safe navigation even in challenging environments.**

I. Introduction

The advancement of autonomous driving technology represents a significant milestone in automotive innovation, capturing global interest and promising to redefine transportation paradigms. Central to this technological revolution is the sophisticated use of sensor-based navigation systems and advanced decision-making algorithms. Among the various approaches, Tesla's camera-centric strategy emphasizes reliance on visual data. However, alternative methodologies incorporating LiDAR technology or a fusion of camera and LiDAR data offer promising avenues to enhance the robustness and reliability of autonomous driving systems.

The synergy between camera and LiDAR technologies is particularly compelling, offering a complementary blend of visual and depth information that can significantly augment the perception capabilities of autonomous vehicles. Cameras provide rich color and texture information, excelling in tasks requiring detailed visual recognition. However, their performance can be compromised under adverse conditions such as low light or direct sunlight. Conversely, LiDAR sensors measure distances using laser light, providing invaluable depth information and maintaining reliability even when cameras falter.

Recognizing the limitations and strengths of each sensor type, this paper proposes a novel method to intelligently determine the most effective use of camera and LiDAR data in real time. Specifically, we introduce a consistency-checking mechanism applied separately to camera-only data and to fused camera-LiDAR data. Our approach is predicated on the assumption that when LiDAR data maintains consistently high quality, it can be used to check the consistency of camera data.

We also employ a feature extraction neural network model inspired by the D2-Net [1] method to assess and validate the reliability of camera data. This consistency-checking process ensures that both individual and fused sensor data maintain high reliability under various conditions, thereby enhancing the overall robustness of the sensor fusion system. By strategically leveraging the unique advantages of each sensor, our approach optimizes their combined potential to enhance autonomous vehicle perception and decision-making.

II. Backgrounds

Advancements in autonomous driving are closely tied to the evolution of 3D detection systems, which are crucial for environmental interpretation and safe navigation. The main methodologies—camera-based,

979-8-3315-2096-0/25/$31.00 ©2025 IEEE

LiDAR-based, and camera-LiDAR fusion—each offer unique advantages and challenges. Camera-based methods capture detailed texture and color information essential for environmental understanding but often struggle with depth perception. LiDAR excels in precise spatial recognition, accurately capturing environmental geometry. Fusion approaches combine the semantic richness of cameras with LiDAR's depth accuracy for enhanced 3D detection.

Benchmarking with datasets like nuScenes [2], KITTI [3], and Waymo [4] has propelled progress in this field. However, research on sensor failures is limited, highlighting the need for adaptive detection systems that maintain safety despite camera or LiDAR malfunctions—critical for reliable autonomous vehicles in real-world scenarios.

Camera-based methods excel in capturing detailed semantic features. Using datasets like KITTI [3], nuScenes [2], and Waymo [4], researchers have developed powerful 3D detection methods based on camera data. Monocular 3D detection is common with KITTI due to its single front camera [5, 6, 7, 8, 9]. Multi-camera setups in nuScenes and Waymo enable Bird's-Eye-View (BEV) 3D detection [10, 11, 12]. Notable advancements include ImVoxelNet [13], which transforms image features into a 3D voxel volume to overcome depth ambiguity, and SMOKE [14], which simplifies monocular 3D object detection by directly estimating 3D bounding box keypoints.

LiDAR-based methods remain prevalent for precise 3D object detection, utilizing depth information. Single-stage detectors like VoxelNet [15], PointPillars [16], and SECOND [17] set benchmarks, while two-stage frameworks incorporate RCNN for refinement [18, 19, 20]. Advancements like CenterPoint [21] focus on object centers to simplify detection and tracking, and CenterFormer integrates transformer networks to enhance detection performance.

Fusion of camera and LiDAR methods, especially through BEV techniques, address limitations of single-sensor approaches by combining camera semantics and LiDAR depth for comprehensive 3D detection. BEVFusion [22] effectively integrates multi-modal data into a unified representation, maintaining spatial and semantic integrity while setting new benchmarks. Our model addresses a gap in BEVFusion research, which often overlooks real-world complexities like camera failures. We propose a dual-model strategy: using BEV fusion under normal conditions and switching to LiDAR-only mode during camera malfunctions, enhancing safety by ensuring reliable data acquisition. Even vision-only systems could benefit from auxiliary LiDAR sensors as failsafes, reinforcing resilience against sensor failures.

Integrating camera and LiDAR data enhances the robustness of autonomous driving systems, ensuring efficient and safe navigation even when one sensor's effectiveness diminishes. By leveraging each sensor's strengths and optimizing their use, our model aims to improve safety and efficiency in autonomous vehicles through reliable sensing mechanisms.

Consistency/Similarity Checking for a Pair of Images

Consistency and similarity checking between image pairs is crucial in computer vision applications like image retrieval, object recognition, and structure-from-motion. Methods include **Siamese Networks** [23, 24], which use twin networks with shared weights to differentiate between similar and dissimilar pairs. **ORB** (Oriented FAST and Rotated BRIEF) [25] is an efficient alternative to SIFT, using binary descriptors that are rotation and scale invariant. The **Scale-Invariant Feature Transform (SIFT)** [26] is a classical method identifying local features invariant to scale and rotation. Recent deep learning-based methods like **SuperPoint** [27] and **D2-Net** [1] jointly learn keypoints and descriptors in an end-to-end manner, offering improved robustness and accuracy. The choice of method depends on specific requirements regarding accuracy, computational efficiency, and robustness to transformations.

III. Data Augmentation/Preparation Methods

In our study, we utilized the nuScenes [2] dataset, renowned for its comprehensive coverage in autonomous driving scenarios. This dataset provides a rich collection of real-world data, encompassing a diverse range of driving conditions and urban landscapes.

A. Camera data consistency detection on Camera-Lidar fusion data

We utilized the nuScenes dataset [2], renowned for its comprehensive coverage of autonomous driving scenarios. Inspired by Yu et al. [28], we defined our data augmentation methods to simulate realistic challenges not typically represented in standard datasets.

We implemented augmentation techniques that simulate these adversities. Specifically, we introduced black blocks into camera images to mimic partial obstructions or camera failures, and added bright spots

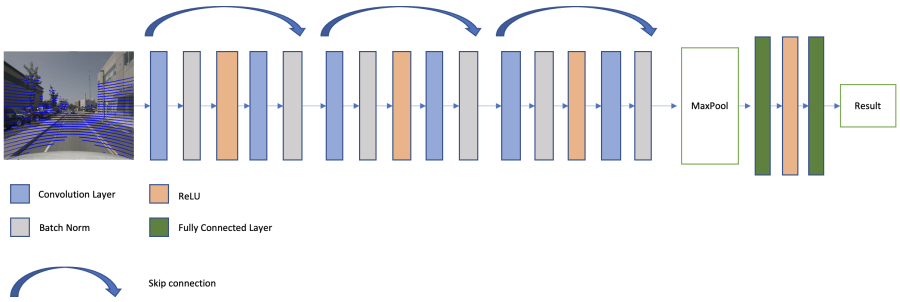

Fig. 1: The Architecture of the Camera Data Checking Model (DetectNet): A Simple Yet Efficient Approach based on camera-lidar fusion.

and global darkening to replicate glare and low-light conditions. By varying the size, intensity, and position of these augmentations, the model is trained to maintain performance even when critical visual information is compromised.

B. Camera-Only Data Consistency Checking

We implemented a camera-only data consistency checking approach using the nuScenes dataset by creating image pairs based on temporal proximity.

We generated two types of pairs: consecutive pairs from the same scene with adjacent timestamps (labeled as consistent) and non-consecutive pairs from different scenes (labeled as inconsistent). This setup ensures that consecutive pairs reflect real-world temporal continuity, while non-consecutive pairs introduce variability and simulate non-continuous driving conditions.

IV. Methodology

Our research methodology leverage the synergy between camera and LiDAR data to ensure data consistency and optimize 3D detection in autonomous driving systems based on our Camera-Lidar data fusion. Also, we developed a comprehensive approach to simulate and detect camera data inconsistencies based on the camera-only dataset, which is crucial for maintaining the reliability of autonomous vehicles. Our methodology consists of feature extraction using the deep learning feature extraction model, followed by feature matching and consistency determination using RANSAC filtering.

A. Camera data consistency detection on Camera-Lidar fusion data

For the task of detecting camera status, we designed a simple yet efficient neural network DetectNet, and we assumed a situation when LiDAR data is valid. This camera status is binary-labeled, with '1' indicating normal camera operation and '0' representing a failure state. We benchmarked our model against well-established neural networks, including ResNet18 by He et al. [29], GoogLeNet by Szegedy et al. [30], and VGG-16 by Simonyan et al. [31] The results indicated that, except VGG-16, the performance was consistent with expectations. The unique aspect of VGG-16's architecture is its use of multiple 3×3 convolutional filters in succession, allowing for a deep network with fewer parameters. However, this series of small filters may sometimes lead to a gradual loss of spatial hierarchies, potentially causing a performance discrepancy in all models that achieved satisfactory performance. Our custom neural network model stood out by offering significantly faster inference times and achieving 94.7% accuracy on the test dataset. This attribute is particularly advantageous for autonomous driving applications, where timely decision-making is critical. Consequently, our model presents a more suitable option for real-time camera consistency checks in the context of autonomous vehicle sensor fusion, balancing both accuracy and computational efficiency.

B. Camera data consistency detection on Camera-only data

For the camera consistency checking based on camera-only data, we deployed three methods, with the deep learning-based approach achieving the best results. Initially, we experimented with ORB (Oriented FAST and Rotated BRIEF) and a Siamese network with a ResNet backbone, but these methods failed to reliably identify pairs with inconsistent data. Consequently, we opted for a deep learning method to extract features.

We employ the D2-Net model [1] to extract features and check the consistency of image pairs. The D2-Net model leverages a single Convolutional Neural Network (CNN) that serves dual purposes as a dense feature descriptor and a feature detector. This method, known as describe-and-detect, postpones the detection stage to leverage more stable high-level features, resulting in robust keypoint detection even under challenging conditions.

V. Experiments and Evaluation

A. Camera-Only Data Consistency Checking

a) Feature Extraction:: We employed D2-Net, a convolutional neural network designed for robust feature extraction. The feature extraction process involves the following steps:

b) Feature Matching and Consistency Determination:: The extracted features are matched using a nearest-neighbor approach with cross-checking. RANSAC filtering is then applied to these matches to identify inliers, which represent consistent feature correspondences between the two images.

The consistency of an image pair is determined based on the number of inliers. The pair is considered consistent if the number of inliers exceeds a predefined threshold. Otherwise, it is labeled as inconsistent.

The accuracy of our camera-only consistency-checking model was evaluated based on its performance on the prepared dataset. The model achieved high accuracy in distinguishing consistent from inconsistent image pairs. Detailed results are showed in Table II.

B. Camera Data Consistency Checking on Camera-LiDAR Fusion Data

We evaluate the model's accuracy using the cross-entropy loss function, following Mao et al. [32], which is suitable for binary classification tasks. The cross-entropy loss measures the difference between the predicted probabilities and the actual labels, encouraging the model to make accurate predictions by penalizing divergence. By minimizing this loss function with a learning rate of 0.001, we refine the model's parameters to enhance its ability to accurately assess camera data consistency, which is crucial for the reliability of autonomous driving applications.

VI. Conclusion

The research presents an advancement in detecting camera failures for autonomous driving using a camera-lidar fusion framework. We also developed a deep learning feature extraction method to check the consistency of camera-only data and a deep learning method to check the consistency of camera data based on camera-lidar fusion, which was tested on the nuScenes dataset.

Our new data augmentation technique, combined with the DetectNet model, has proven to be highly effective in addressing the camera-lidar consistency checking problem, achieving an accuracy of over 95% and demonstrating high efficiency. Additionally, our deep learning model, which extracts and maps features from image pairs using advanced feature extraction techniques, achieved perfect accuracy in the camera-only consistency checking problem.

VII. Acknowledgment

This research was partly supported by NSF award (ID: 2010366) to Dr. Fang and Dr. Wang.

References

[1] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler, "D2-Net: A Trainable CNN for Joint Detection and Description of Local Features," *CVPR*, 2019.

[2] 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," *arXiv preprint arXiv:1903.11027*, 2019.

[3] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets Robotics: The KITTI Dataset," *International Journal of Robotics Research (IJRR)*, 2013.

[4] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, Y. Zhang, J. Shlens, Z. Chen, and D. Anguelov, "Scalability in perception for autonomous driving: Waymo open dataset," *CVPR*, 2020.

[5] Z. Liu, D. Zhou, F. Lu, J. Fang, and L. Zhang, "Autoshape: Real-time shape-aware monocular 3D object detection," *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[6] Y. Lu, X. Ma, L. Yang, T. Zhang, Y. Liu, Q. Chu, J. Yan, and W. Ouyang, "Geometry uncertainty projection network for monocular 3D object detection," *IEEE International Conference on Computer Vision (ICCV)*, 2021.

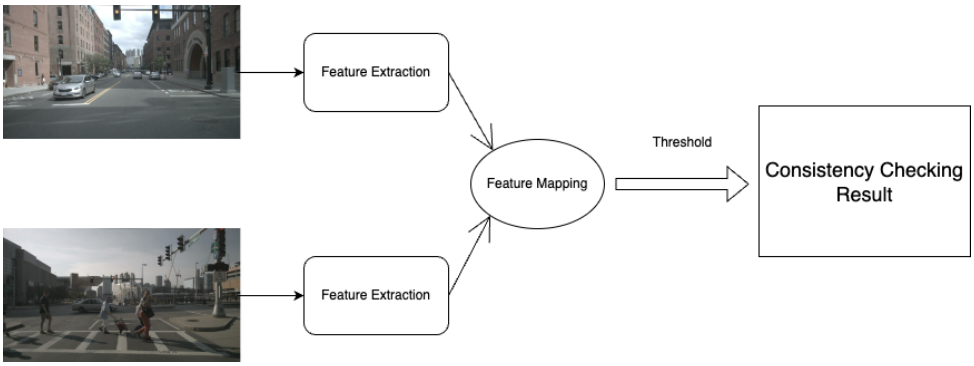

Fig. 2: This framework shows the key idea of consistency checking based on camera-only data.

Models + Backbone	**Distance Method**	**Threshold**	**Accuracy**
Siamese Network + ResNet18	Manhattan Distance	Distance $<$ 1.0	51.12%
Siamese Network + ResNet50	Manhattan Distance	Distance $<$ 1.0	53.12%
Siamese Network + ResNet18	Euclidean Distance	Distance $<$ 1.0	47.8%
Siamese Network + ResNet50	Euclidean Distance	Distance $<$ 1.0	46.35%
ORB (Oriented FAST and Rotated BRIEF)	Nan	200	42.3%
ORB (Oriented FAST and Rotated BRIEF)	Nan	100	46.4%
D2-NET + VGG16	Nan	50	94%
D2-NET + ResNet50	Nan	200	100%

TABLE I: Consistency Checking Results

Fig. 3: This figure shows an example of consistent camera data.

Model	Accu on Training	Accu on Valid	Accu on Test
DetectNet	0.961	0.955	0.947
Resnet18	0.962	0.957	0.968
GoogleNet	0.954	0.965	0.949
VGG-16	0.767	0.757	0.760

TABLE II: This table compares the performance of our DetectNet model with that of VGG-16, ResNet18, and GoogleNet in terms of accuracy on training, validation, and test datasets for camera data consistency checks.

[7] A. Kumar, G. Brazil, and X. Liu, "Groomed-nms: Grouped mathematically differentiable nms for monocular 3D object detection," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[8] L. Wang, L. Du, X. Ye, Y. Fu, G. Guo, X. Xue, J. Feng, and L. Zhang, "Depth-conditioned dynamic message propagation for monocular 3D object detection," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[9] L. Wang, L. Zhang, Y. Zhu, Z. Zhang, T. He, M. Li, and X. Xue, "Progressive coordinate transforms for monocular 3D object detection," *Neural Information Processing Systems (NeurIPS)*, 2021.

[10] T. Wang, X. Zhu, J. Pang, and D. Lin, "Probabilistic and geometric depth: Detecting objects in perspective," *Conference on Robot Learning (CoRL)*, 2022.

[11] T. Wang, X. Zhu, J. Pang, and D. Lin, "Fcos3d: Fully convolutional one-stage monocular 3D object detection," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[12] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "Detr3d: 3D object detection from multi-view images via 3D-to-2D queries," *Conference on Robot Learning (CoRL)*, 2022.

[13] D. Rukhovich, A. Vorontsova, and A. Konushin, "ImVoxelNet: Image to Voxels Projection for Monocular and Multi-View General-Purpose 3D Object Detection," *Neural Information Processing Systems (NeurIPS)*, 2022.

[14] Z. Liu, Z. Wu, and R. Tóth, "SMOKE: Single-Stage Monocular 3D Object Detection via Keypoint Estimation," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

[15] Y. Zhou and O. Tuzel, "VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection," *CVPR*, 2018.

[16] A. H. Lang, S. Vora, H. Caesar, L. Zhou, and J. Yang, "PointPillars: Fast Encoders for Object Detection from Point Clouds," *CVPR*, 2019.

[17] Y. Yan, Y. Mao, and B. Li, "SECOND: Sparsely Embedded Convolutional Detection," *Sensors*, 2018.

[18] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li, "PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection," *CVPR*, 2020.

[19] S. Shi, Z. Wang, J. Shi, X. Wang, and H. Li, "From Points to Parts: 3D Object Detection from Point Cloud with Part-aware and Part-aggregation Network," *TPAMI*, 2020.

[20] Y. Chen, S. Liu, X. Shen, and J. Jia, "Fast Point R-CNN," *ICCV*, 2019.

[21] T. Yin, X. Zhou, and P. Krähenbühl, "Center-Former: Center-based Transformer for 3D Object Detection," *CVPR*, 2021.

[22] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. Rus, and S. Han, "BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation," *Neural Information Processing Systems (NeurIPS)*, 2022.

[23] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a 'Siamese' time delay neural network," *International Journal of Pattern Recognition and Artificial Intelligence*, 1993.

[24] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," *ICML Deep Learning Workshop*, 2015.

[25] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," *International conference on computer vision*, 2011.

[26] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, 2004.

[27] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," *CVPR Workshops*, 2018.

[28] K. Yu, T. Tang, H. Xie, Z. Lin, Z. Wu, Z. Xia, T. Liang, H. Sun, J. Deng, D. Hao, Y. Wang, X. Liang, and B. Wang, "Benchmarking the Robustness of LiDAR-Camera Fusion for 3D Object Detection," *arXiv:2205.14951 [cs.CV]*, 2022.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Benchmarking the Robustness of LiDAR-Camera Fusion for 3D Object Detection," *arXiv:1512.03385 [cs.CV]*, 2015.

[30] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *arXiv:1409.4842 [cs.CV]*, 2014.

[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv:1409.1556 [cs.CV]*, 2014.

[32] A. Mao, M. Mohri, and Y. Zhong, "Cross-Entropy Loss Functions: Theoretical Analysis and Applications," *arXiv:2304.07288 [cs.LG]*, 2023.