Federated Monocular 3D Object Detection for Autonomous Driving

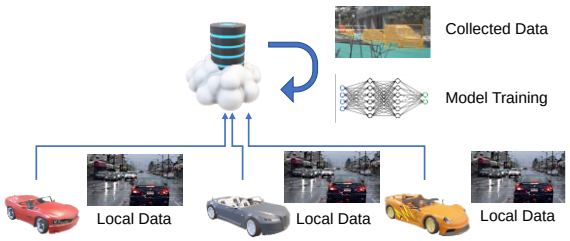
Fangyuan Chi, Yixiao Wang, Panos Nasiopoulos, Victor C.M. Leung, Mahsa T. Pourazad
Department of Electrical and Computer Engineering
The University of British Columbia
Vancouver, BC, Canada
{fangchi, yixiaow, panosn, vleung, pourazad}@ece.ubc.ca

***Abstract*—In this paper, we propose and implement a novel method for 3D object detection in autonomous driving by applying federated mechanism to a monocular camera-based network. Our approach has several advantages over traditional 3D object detection methods that rely on LiDAR or other sensors, as it is more cost-effective and can be more easily integrated into existing autonomous driving systems. We use a federated learning framework, which allows us to train the model on a large amount of data covering a variety of scenarios without having to share the raw data with a central server. This allows us to reduce transmission bandwidth requirements and preserve the privacy of the data contributors, while still achieving high accuracy in 3D object detection. In our experiments, we evaluate our method on a variety of challenging real-world driving scenarios and show that it is able to accurately detect objects in 3D from a monocular camera view. Our results demonstrate the effectiveness of our approach and show its potential for use in autonomous driving systems.**

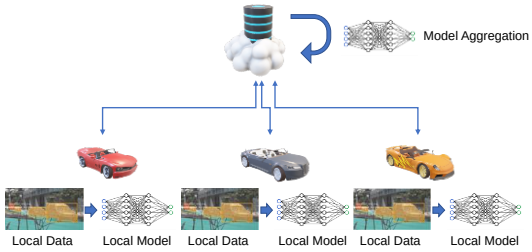
Keywords-monocular, 3D object detection, federated learning, autonomous driving.

I. INTRODUCTION

Safety remains the primary concern when people talk about autonomous driving. Autonomous vehicles are empowered by various Deep Learning (DL) models (i.e., perception, tracking, prediction, etc.), but these models are trained with either simulated data or controlled driving scenarios, with most autonomous vehicles still being tested in enclosed facility environments. It is, therefore, difficult to evaluate how they will perform in real-world driving scenarios, especially when the controlled environment is coupled with numerous unpredictable corners, emergencies, and occlusions. Human drivers gain experience over time, first watching the parents or others driving, then through driving schools and, finally, driving and improving their skills every year. Autonomous vehicles should do the same. Besides training deep learning models that enable autonomous driving in labs, autonomous vehicles should continuously learn from different driving scenarios of their own or others experience. Federated Learning (FL) is a promising approach that may address this problem. Instead of having autonomous vehicles to upload their perception data to the cloud to perform centralized training, as shown in Figure 1(a), FL allows autonomous vehicles to first train their local models with local collected data and share with each other their own experience through their DL models instead of sharing collected data, as shown in Figure 1(b). Recent works investigate if autonomous driving could benefit from FL. Some of them [1][2] are designing system architectures to ensure the efficiency of when and which vehicles should participate in the federation process. Other approaches are trying to address the problem that data collected locally are non-Independent and Identically Distributed (non-IID) [3][4]. Although existing methods demonstrate that FL could improve the accuracy in object detection, they have only been evaluated with 2D object detection [5] or 3D object detection using LiDAR [6]. However, it is not sufficient to use these two types of sensor data when evaluating methods for autonomous driving. More specifically, 2D object detection cannot output depth information or is hard to predict the distance of target objects, while LiDAR is expensive and is not commonly supported by autonomous vehicles, with the industry recently following a vision-based trend for autonomous driving [7][8]. In this paper, we investigate and verify that the performance of 3D object detection could benefit from leveraging federated learning with 3D image data collected by monocular cameras.



(a) Centralized Monocular Approach



(b) Federated Monocular Approach

Figure 1. (a) Centralized approach which upload the local images from all vehicles to the central server, then train and release the global model, (b) Federated approach, in which each vehicle trains their own model with the local data, then the local models will be uploaded, aggreated and released.

 ISBN: ISBNFILL

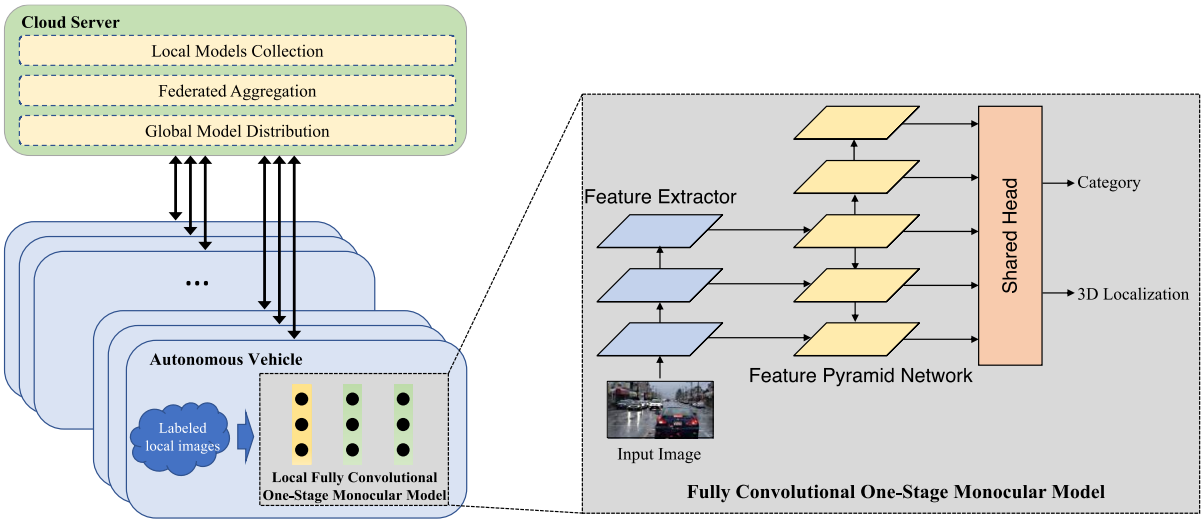


Figure 2. Overall architecture of the federated monocular 3D object detection approach.

The rest of the paper is organized as follows. Section II gives an overview of related work, while Section III presents our proposed approach. In Section IV, we evaluate and analyze the results. Section V concludes the paper.

II. Overview of Related Work

A. 3D Object Detection in Autonomous Driving

Object detection is a popular research topic in autonomous driving. Different approaches are proposed leveraging different kinds of sensors. For example, in [9] the RGB image captured by camera and point cloud obtained by LiDAR are fused together for 3D object detection. Evaluation results show that the detection accuracy increases since data from different modalities provide complimentary features (i.e., images provide semantic information while point cloud provides depth information to construct 3D surroundings). In [10], data obtained from LiDAR are used for 3D object detection. This approach considers long-range interactions among detection candidates. In [11], RGB-D images captured by monocular camera are used for 3D object detection. This vision-based approach is simpler, cost-efficient, and more practical compared to multi-modality-based approaches.

B. Federated Learning in Autonomous Driving

Federated learning is a machine learning approach that allows multiple participants to train a shared model without sharing their raw data. This is particularly useful in the context of autonomous driving, where data from individual vehicles may be sensitive or proprietary. With federated learning, each vehicle can train a local model on its own data, and then share the model updates with a central server. The server can then aggregate the updates and use them to improve a shared global model, without ever having access to the raw data. This approach has several potential benefits for autonomous driving. For example, it allows vehicles to learn from each other without sharing sensitive data, and can enable the development of more robust and accurate models by leveraging data from a larger and more diverse set of vehicles [1]-[6]. Additionally, federated learning can enable real-time updates to the global model, allowing vehicles to quickly adapt to changing conditions and improve their performance over time. Overall, federated learning has the potential to play a significant role in the development of autonomous driving systems.

III. Our Proposed Method

In this paper, we propose a federated monocular 3D object detection approach for autonomous driving. The overall architecture of our method is illustrated in Figure 2. The left part of the figure shows the distributed federated learning mechanism, while the right part shows the local monocular model on each vehicle, which is trained to make predictions for 3D object detection.

A. Federated Learning-based Collaboration

The federated learning mechanism adopted in our approach includes the following 3 steps. 1) First, each vehicle trains a local monocular model on its own. This allows it to learn from its own data without sharing it with other vehicles or a central server. 2) After training the local models, each vehicle shares the model updates (e.g., weights, biases) with the central server. The server can then aggregate all these updates and use them to improve the global model. 3) Once the global model has been updated, the central server distributes the updated model to all the vehicles. Each vehicle can then use the updated global model to improve its own local model and continue the training process. This 3-step process can be repeated as necessary to continue improving the performance of the global model and enable vehicles to learn from each other. Over time, the global model should become more accurate and robust, allowing vehicles to make better decisions and improve their individual performance in real world conditions. Note that, during the federated training

 ISBN: ISBNFILL

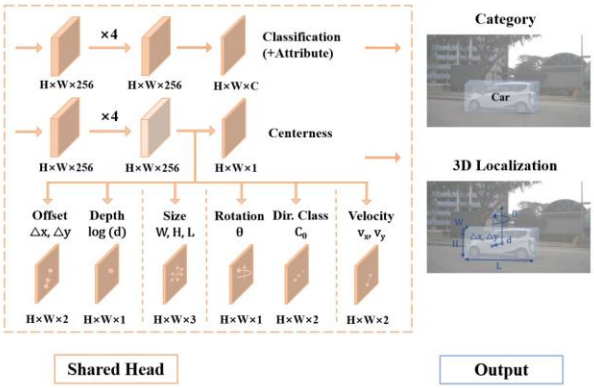


Figure 3. Illustration of the shared head in the local monocular 3D object detection model.

process, the images captured on each vehicle are not sent to the central server, keeping possible sensitive information on the local vehicle, thus eliminating privacy concern issues.

B. Local Monocular 3D Object Detection Model

We employ a fully convolutional one-stage monocular 3D object detection model for detecting objects in 3D from a single camera view [11]. This is an object detection approach that uses a Convolutional Neural Network (CNN) to predict the 3D bounding boxes and class probabilities of objects in an image.

One key advantage of this approach is that it is fully convolutional, meaning that the CNN can operate on input images of any size, and produce output predictions for each pixel in the image. This allows the model to be used on images of varying resolutions, without the need for manual resizing or cropping. Additionally, this approach is a one-stage method, meaning that it uses a single neural network to make all of its predictions. This makes the model more efficient and faster to run, as it does not require multiple stages of processing or separate networks for different tasks.

In the local monocular 3D object detection model, the ResNet101 [12] is employed as the feature extractor and the Feature Pyramid Network (FPN) [13] as the neck. The ResNet101 network is a well-known and widely used architecture for image classification and object detection tasks. It is a deep CNN that is composed of multiple convolutional layers, residual blocks, and pooling layers, and is designed to be highly efficient and accurate. By using ResNet101 as the feature extractor in our model, we can take advantage of its proven performance and efficiency, and extract high-quality features from the input images. The FPN is a network architecture that is commonly used in object detection tasks to improve the model's ability to detect objects at different scales. It is composed of a pyramid of feature maps, with each level of the pyramid representing features at a different scale. By using FPN as the neck in our model, we can improve the model's ability to detect objects such as pedestrians and cars at different distances from the camera.

The FPN neck is followed by a shared head, shown in Figure 3. Using a shared head to output the class of objects and 3D bounding boxes in a fully convolutional one-stage monocular 3D object detection model can have several benefits. First, a shared head allows the model to make predictions for both the class of objects and the 3D bounding box in a single pass, which can make the model more efficient and faster to run. This is particularly useful in real-time applications such as autonomous driving, where it is important to make predictions quickly and accurately. Second, a shared head can improve the model's overall performance, as it allows the CNN to learn features that are relevant for both tasks simultaneously. This not only can lead to more accurate predictions but also to better generalization to new data. Finally, a shared head can simplify the model's architecture, making it easier to train and optimize. This can save time and resources, and can make the model more portable and easier to integrate into different applications.

IV. Experimental Results and Discussion

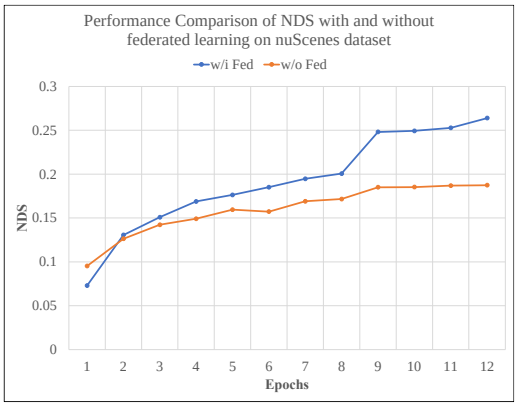
We used the nuScenes dataset to evaluate the performance of our federated monocular 3D object detection method. The nuScenes dataset is a large-scale dataset of annotated images and point clouds captured in real-world driving scenarios [14]. It contains a rich variety of data, including different environments, weather conditions, and vehicle types, making it an ideal testbed for evaluating our method. To perform our evaluation, we split the nuScenes dataset into 10 parts, each representing data from a different vehicle.

We then train our local monocular models on each of these vehicle datasets (10% of the whole dataset), both with and without federated learning. This allows us to compare the performance of our method with and without federated learning. We train all the models for 12 epochs at a batch size of 4 on a NVIDIA Tesla V100L graphic card. The learning rate is set to $5e^{-3}$ and is halved in both the 8^{th} and 11^{th} epoch. For our approach, the trained model weights from the 10 vehicles are aggregated in an average manner in every epoch.

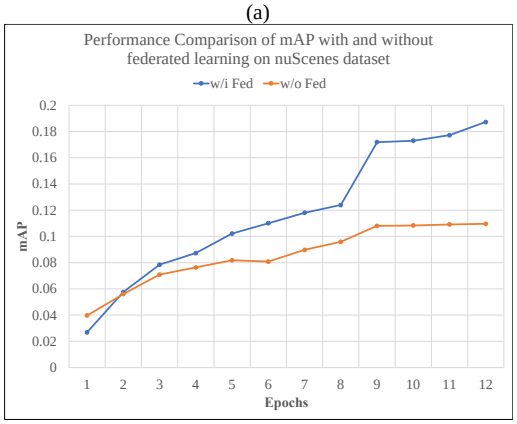
After training our federated monocular 3D object detection model, we use the mean Average Precision (mAP) and NuScenes Detection Score (NDS) metrics to evaluate its performance. These metrics are commonly used to evaluate object detection algorithms, and allow us to compare our method to other state-of-the-art approaches. The mAP metric measures the average precision of the model across all classes and all thresholds. It is calculated by averaging the precision of the model at different recall levels, and is a useful metric for comparing the overall performance of different object detection models. The NDS metric is specific to the nuScenes dataset, and measures the overall performance of the model in terms of both precision and recall. It is calculated as the harmonic mean of the average precision and average recall of the model, and is a useful metric for evaluating object detection models on the nuScenes dataset.

Figure 4 (a) shows the performance comparison of our method with and without federated learning on the nuScenes dataset in NDS against epochs, while Figure 4 (b) shows the same comparison in mAP vs epochs. We observe that the NDS of our method is significantly higher when using federated learning, reaching 41.18%. This demonstrates that our method is able to improve both the precision and recall of its predictions, resulting in more complete and accurate

 ISBN: ISBNFILL



(a)



(b)

Figure 4. The validation results comparison at each training epoch on two metrics, (a) NDS, (b) mAP

detections of objects in the scene. Similarly, we observe that mAP of our method is 70% higher when using federated learning, indicating that our method is able to make more accurate predictions when using data from multiple vehicles, as opposed to training only on local data from a single vehicle. The performance is summarized in Table I.

Overall, our evaluation results show that the adopted federated learning mechanism is able to significantly improve the prediction performance of our federated monocular 3D object detection approach, leading to higher mAP and NDS scores compared to training only with local data. This demonstrates the effectiveness of our method and its potential for use in autonomous driving systems.

V. CONCLUSION

In conclusion, this paper has presented a novel method for 3D object detection in autonomous driving using only a monocular camera. Our approach uses federated learning to train a deep neural network that is able to detect objects in 3D from a single camera view, and has several advantages over traditional methods that rely on LiDAR or other sensors. We have evaluated our method on a variety of challenging real-world driving scenarios and showed that it is able to accurately detect objects in 3D from a monocular camera view. These results demonstrate the effectiveness of our approach and suggest its potential for use in autonomous driving systems.

TABLE I. COMPARISON OF NDS AND MAP METRICS BETWEEN THE BASELINE AND OUR PROPOSED FEDERATED APPROACH

Method	Backbone	Data Ratio	NDS	mAP
Baseline	ResNet101	10%	0.187	0.110
Ours	ResNet101	10%	**0.264 (+41.18%)**	**0.187 (+70%)**

REFERENCES

[1] A. Nguyen et al., "Deep Federated Learning for Autonomous Driving," 2022 IEEE Intelligent Vehicles Symposium (IV), 2022, pp. 1824-1830.

[2] S. Wang et al. Federated Deep Learning Meets Autonomous Vehicle Perception: Design and Verification[J]. arXiv preprint arXiv:2206.01748, 2022

[3] Q. Li, B. He, and D. Song, "Model-contrastive federated learning" Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021, pp. 10713-10722.

[4] T. Zeng, O. Semiariy, M. Chen, W. Saad and M. Bennis, "Federated Learning on the Road Autonomous Controller Design for Connected and Autonomous Vehicles," IEEE Transactions on Wireless Communications, 2022, vol. 21, no. 12, pp. 10407-10423

[5] D. Jallepalli, N. C. Ravikumar, P. V. Badarinath, S. Uchil and M. A. Suresh, "Federated Learning for Object Detection in Autonomous Vehicles," IEEE Seventh International Conference on Big Data Computing Service and Applications (BigDataService), 2021, pp. 107-114.

[6] B. Luca, S. Stefano, B. Mattia, and N. Monica, "Decentralized federated learning for extended sensing in 6G connected vehicles", Vehicular Communications, 2022, vol. 33, pp.100396, ISSN 2214-2096

[7] R. Trabelsi, R. Khemmar, B. Decoux, J. Y. Ertaud, and R. Butteau, "Recent advances in vision-based on-road behaviors understanding: a critical survey". Sensors, 2022, vol. 22, no. 7, pp. 2654

[8] E. Zablocki, H. Ben-Younes, P. Pérez, and M. Cord. "Explainability of deep vision-based autonomous driving systems: Review and challenges". International Journal of Computer Vision, 2022, vol. 130, no. 10, pp. 2425-2452.

[9] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3D Object Detection Network for Autonomous Driving," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, Jul. 2017, pp. 6526–6534.

[10] C. He, H. Zeng, J. Huang, X. Hua, and L. Zhang, "Structure aware single-stage 3d object detection from point cloud" Proc. IEEE/CVF conference on computer vision and pattern recognition. 2020, pp. 11873-11882.

[11] T. Wang, X. Zhu, J. Pang, and L. Dahua, "Fcos3d: Fully convolutional one-stage monocular 3d object detection" in Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021, pp. 913-922.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[13] T. Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, Jul. 2017, pp. 936–944.

[14] H. Caesar et al., 'nuScenes: A Multimodal Dataset for Autonomous Driving', IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, Jun. 2020, pp. 11618–11628.

 ISBN: ISBNFILL