# CoFF: Cooperative Spatial Feature Fusion for 3D Object Detection on Autonomous Vehicles

Jingda Guo, Dominic Carrillo, Sihai Tang, Qi Chen, Qing Yang, *Senior Member, IEEE,* Song Fu, *Senior Member, IEEE,* Xi Wang *Member, IEEE,* Nannan Wang *Member, IEEE,* Paparao Palacharla *Member, IEEE*

***Abstract*—To reduce the amount of transmitted data, feature map based fusion is recently proposed as a practical solution to cooperative 3D object detection by autonomous vehicles. The precision of object detection, however, may require significant improvement, especially for objects that are far away or occluded. To address this critical issue for the safety of autonomous vehicles and human beings, we propose a cooperative spatial feature fusion (CoFF) method for autonomous vehicles to effectively fuse feature maps for achieving a higher 3D object detection performance. Specially, CoFF differentiates weights among feature maps for a more guided fusion, based on how much new semantic information is provided by the received feature maps. It also enhances the inconspicuous features corresponding to far/occluded objects to improve their detection precision. Experimental results show that CoFF achieves a significant improvement in terms of both detection precision and effective detection range for autonomous vehicles, compared to previous feature fusion solutions.**

***Index Terms*—Autonomous vehicles, cooperative perception, 3D object detection, feature fusion, feature enhancement.**

## I. Introduction

An autonomous vehicle (AV) relies on its perception system to sense the surroundings and makes driving decisions accordingly. However, the sensors equipped on AV are typically non-line-of-sight, whose effective sensing range could be significantly reduced on crowded roads, due to blockages and/or occlusions. Therefore, it is crucial to connect AVs and allow them to exchange sensor data to facilitate precise cooperative perception, thus improving driving safety. A major challenge for cooperative perception on connected and autonomous vehicles (CAVs) is how to effectively merge sensor data received from different AVs to obtain a precise and comprehensive perception.

Due to the huge volume of raw sensor data, it is practically infeasible to exchange raw data among vehicles, which would cause severe bottlenecks in existing network infrastructures. To reduce network traffic, a feature map based data sharing mechanism is proposed for 3D object detection on autonomous vehicles [1]. Feature maps are the intermediate results produced by a Convolutional Neural Network (CNN). In [1], feature maps generated on different vehicles are combined to yield a cooperative object detection. However, we find that its feature fusion mechanism can be further improved, by considering the volume of new semantic information contained in the to-be-fused feature maps and enhancing weak feature on feature maps. In this article, we aim to achieve an enhanced object detection performance by fusing feature maps in an intelligent manner.

Unlike raw sensor data, feature maps are hard to interpret, which increases the difficulty in designing an effective fusion mechanism for cooperative 3D object detection. State-of-the-art feature maps fusion solutions overlook the semantic information difference between to-be-fused feature maps, lead to the adverse effect caused by receivers' feature maps, and affect the detection performance of cooperative perception. To tackle this challenge, we investigate how the volume of new semantic information provided by a receiver's feature map influences its importance on fusion. We call this approach "cooperative spatial feature fusion". We hypothesize that feature maps produced by a distant vehicle can significantly improve the object detection on the current vehicle, particularly for recognizing distant objects. Moreover, we find the weak feature of far/occluded objects can be enhanced for a better detection performance after fusion, even they are hard to be detected by current approaches due to far distance or occlusion. Therefore, a better object detection performance is expected if feature maps are fused in a more convincing way, considering enhancement on features. Towards this end, we propose a novel cooperative spatial feature fusion mechanism for CAVs to fuse feature maps and achieve accurate 3D object detection effectively.

### *A. Main Challenges*

In designing the spatial feature fusion method, we need to conquer two major technical challenges. The first challenge is how to identify and reduce the negative effects on object detection caused by mistakenly fused feature maps generated by different vehicles. This problem was overlooked in the existing work [1] as it does not consider how feature maps affect each other when they are fused in a wrong way. The underlying fusing function adopted by the previous work is *maxout* which selects the features with larger values in the fusion process. The method seems reasonable as it keeps the most distinctive features while suppressing non-distinctive ones; however, it might omit important features received from other vehicles, which could have significantly improved the current vehicle's object detection performance if used correctly. In other words, the feature maps generated by multiple vehicles should be treated differently, instead of equally as was the case in the previous work [1].

Jingda Guo is the corresponding author.

J. Guo, D. Carrillo, S. Tang, Q. Chen, Q. Yang, and S. Fu are with the Department of Computer Science and Engineering, University of North Texas, Deton, TX, 76207 USA. e-mail: (JingdaGuo@my.unt.edu, DominicCarrillo@my.unt.edu, sihaitang@my.unt.edu, qichen@my.unt.edu, qing.yang@unt.edu, song.fu@unt.edu).

X. Wang, N. Wang and P. Palacharla are with Fujitsu.

Manuscript received August 29, 2020; revised December 24, 2020; accepted January 06, 2021.

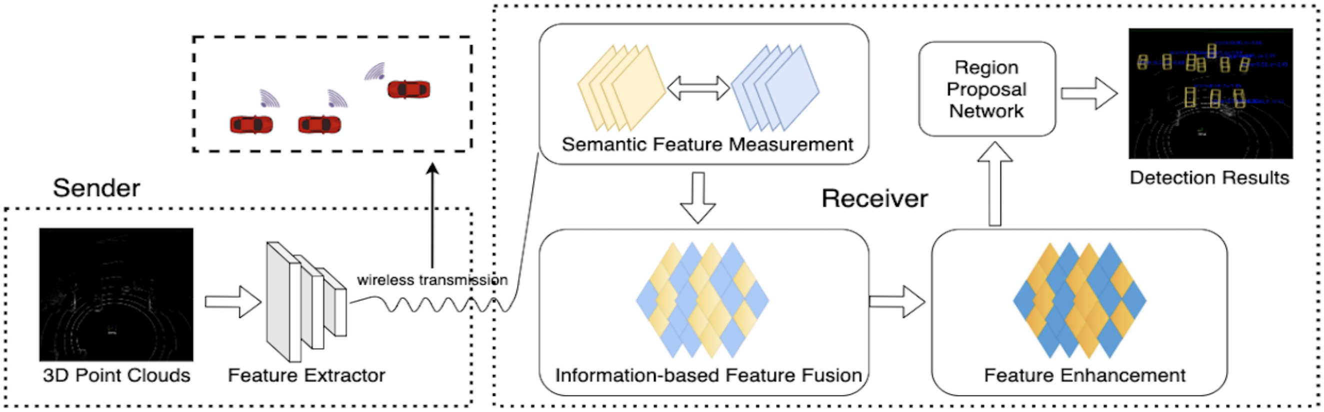


Figure 1: Overview of CoFF

The second challenge lies in the difficulty of detecting distant or occluded objects. This is a problem not just for cooperative object detection, but a common issue for many 3D object detection approaches for autonomous vehicles. When feature maps are fused and more information is considered, we can detect objects which are otherwise hard to detect due to insufficient information from individual sensor data. This is because feature maps generated by different vehicles complement each other, and if fused properly, they can offer a more comprehensive representation of objects.

### B. Proposed Solution

In this paper, we propose our cooperative perception solution CoFF (Cooperative spatial Feature Fusion) to 3D object detection on autonomous vehicles to conquer both above-mentioned issues that exist on state-of-the-art cooperative perception solutions on CAV. Specifically, we design a dynamic method, called information based spatial feature fusion, to weight the proportion of feature maps on fusion, based on how much new information brings from coming feature maps. For objects that are hard to detect, caused either by occlusion or far distance, we propose a feature enhancement approach to enhance their feature on fused feature maps and improve detection performance. The extra computation and communication overhead of CoFF are negligible for current hardware equipped on autonomous vehicles, which guarantees our CoFF remains a real-time 3D cooperative perception approach on CAV.

We refer the reader to Fig. 1 for the overview of CoFF. 3D point cloud data collected by sender AVs is first processed by feature extractors to generate spatial-aware feature maps for transmission. Then source AV transmit its feature maps to nearby AVs. After a receiver AV receives the feature maps, it then evaluates the value of the feature maps, measuring by how much new semantic information can be provided compared to its own feature maps. Therefore, the weight of information-based feature fusion is determined, and feature maps with more new semantic information have a larger weight on fusion. The resulting feature maps are sent to the feature enhancement module, which linearly enhances features on feature maps, for a better detection performance. The last part of CoFF is Region Proposal Network, generating detection results for CoFF, including the classes and coordinates of detected objects.

### C. Contributions

The main contributions of this article can be summarized as follows. First, we propose a new feature fusion approach to cooperative perception on autonomous vehicles, aiming to improve the 3D object detection performance, especially on distant or occluded objects. Our novel idea is to factor in the new semantic information when fusing feature maps from different vehicles, e.g., greater weights are given to the feature maps containing more new semantic information compared with receiver's feature maps. Second, we discover that the numerical values on 3D detection feature maps usually represent the significance of the underlying features on detection. Thus, enhancing features on feature maps and enlarging the difference between the features representing the objects and those representing the background can improve detection performance. Our proposed feature enhancement method linearly enhances feature on feature maps, aiming to increase the values of the features representing objects while keeping the detection results on other parts almost unchanged. Our proposed method is generic and applicable to other applications that involve fusing 3D data/features generated by different sensors/entities, e.g., in Internet of Things environments.

## II. Preliminaries and Background

It has been shown that sharing raw LiDAR (Light Detection and Ranging) data among autonomous vehicles can help 3D object detection on individual vehicles. The basic idea of cooperative perception, Cooper [2], is to fuse LiDAR point cloud data produced by multiple vehicles to cooperatively detect 3D objects. While Cooper [2] provides a means for raw sensor data fusion to improve object detection performance, transmitting raw point cloud data places a heavy burden on vehicle-to-vehicle (V2V) wireless networks. One frame of 64-beam LiDAR data can be as large as 3 MB, and a typical LiDAR can generate as many as 20 frames per second, equivalent to 480 Mbps of network capacity. Therefore, it is difficult to continuously transmit such massive amount

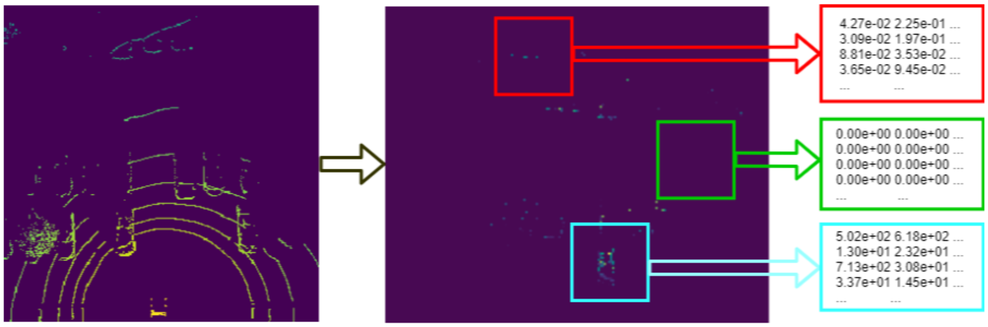


Figure 2: Illustration of a feature map generated by F-Cooper [1] from LiDAR point cloud data. Examples of strong features, weak features, and background features are depicted in blue, red, and green boxes, respectively.

of data over today's wireless networks. As an intermediate result from a CNN-based detection model, feature maps carry important semantic information for object detection and can be considered a substitute for the original sensor data for cooperative object detection.

In this section, we briefly describe feature maps for 3D object detection in Section II-A. We explain feature map based fusion for cooperative perception on AVs in Section II-B and discuss the limitations of existing approaches in Section II-C.

### A. Feature Maps for 3D Object Detection

With the rapid growth and increasing capability of CNN, most state-of-the-art object detection models, such as [3], [4], [5], on autonomous vehicles are CNN-based. Feature maps, as the intermediate representations of CNNs, is the output of feature extractors [6], and contain important semantic information to accomplish object recognizing tasks. Therefore, feature maps provide necessary semantic information for further processing, with only being a fraction of the original data size.

For 3D object detection, many pioneers' works [7], [8], [9] divide 3D space into voxels and generate corresponding location-aware features using 3D convolutions. As a representative solution to CNN-based 3D object detection, VoxelNet [9] becomes the backbone network for many state-of-the-art 3D detectors (e.g., [10], [11]). The feature extractor of F-Cooper [1] leverages the design of VoxelNet, which first divides original 3D point cloud data into thousands of voxels, and build spatial-aware feature maps based on number of points in voxels for detection.

Fig. 2 shows an example of a spatial feature map generated from F-Cooper on LiDAR data. Voxels containing more point clouds typically show prominent features, e.g., those depicted in the blue box in the figure have greater values on average than other parts in the feature map. In contrast, the values enclosed in the red box are smaller than those in the blue box, indicating fewer point cloud data collected from that region. For the voxels containing no points, i.e., they do not provide any useful information for object detection, their corresponding values in the feature map are all zeros, as shown in the green box. Here, we define a feature in the feature map that contains more larger values as a **strong feature**, while a feature with fewer larger values as a **weak feature**. We will use these definitions to compare the importance of two feature maps, which will be crucial for designing the feature fusion mechanism in later discussion.

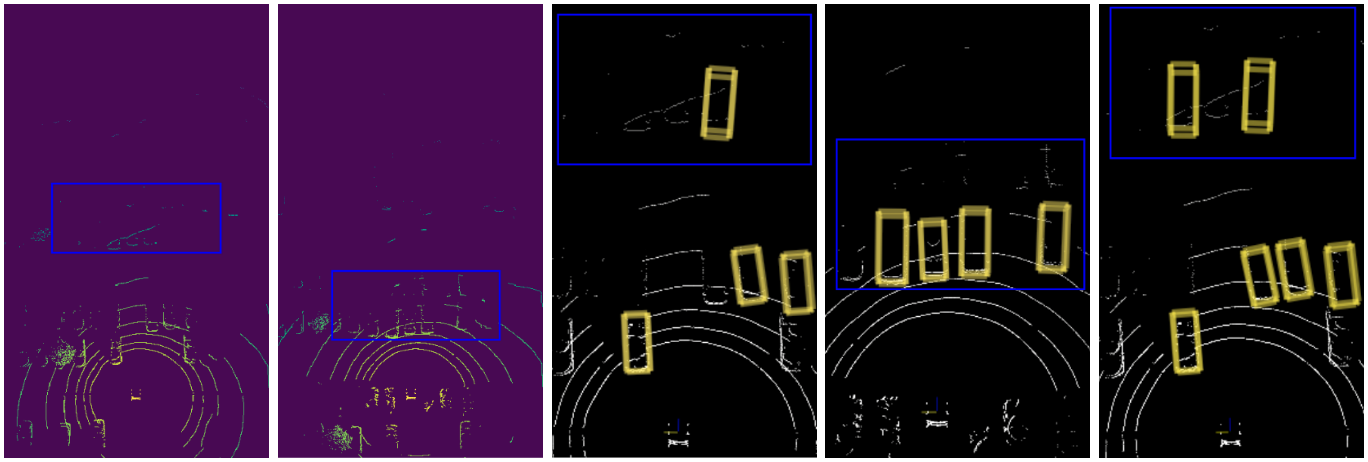
### B. Feature Map Based Fusion

Feature maps can be considered a substitute for the original sensor data; therefore, cooperative perception on CAVs can also be achieved by fusing feature maps. F-Cooper [1] is a state-of-the-art solution that achieves cooperative 3D object detection by sharing feature maps among autonomous vehicles. For an AV that receives a feature map from a nearby AV, it fuses the received feature map with its own feature map by aligning them based on their physical locations. The location information can be obtained from the corresponding point cloud data. The fusion of the feature maps can be viewed as creating a new feature map that contains merged features. Experimental results show that the $maxout$ function [12] helps F-Cooper detect more objects from the fused feature maps, including objects that cannot be detected by individual sender/receiver AV.

### C. Limitations of F-Cooper

There are two major limitations originated from the $maxout$ fusion function in F-Cooper [1]. First, F-Cooper does not consider the importance of individual to-be-fused feature maps. Second, F-Cooper tends to have difficulty in detecting distant or occluded objects.

A illustrative example is shown in Fig. 3, in which a sender vehicle is sharing its sensor data, in the feature maps format, to a receiver vehicle located behind the sender vehicle. For the same region (depicted in the blue box in Fig. 3(a) and (b)), stronger features are more likely to be generated by the sender, and therefore have a better object detection performance for this region, as shown in Fig. 3(c) and (d), due to its physical proximity to the region. For the same region, as it is relatively far from the receiver, as shown in Fig. 3(a), the receiver would generate weak features on its feature map. Due to laser scattering and occlusion, however, some values for that region in the receiver's feature map could be larger than those in the sender's feature map. As F-Cooper [1] treats all feature maps equally on fusion with $maxout$ function, which essentially keeps larger values of two feature maps, partial of the corresponding features provided by the sender will be removed. And weak features from the receiver's feature map will affect the overall detection



(a) 3D data of receiver (b) 3D data of sender (c) Detection results on receiver (d) Detection results on sender (e) F-Cooper's results

Figure 3: Limitations of F-Cooper fusion on detecting far objects. Detection results on far objects of F-Cooper is not as good as those on the sender side, shown in the blue box.

performance on the fused feature map. As shown in Fig. 3(e), two objects that were detected by the sender do not show up in the detection results after fusion. The negative effects of weak features on cooperative object detection become more prominent when occlusions occur more frequently, e.g., in a heavy traffic scenario.

As laser signals are more likely to be blocked by nearby objects or effuse into the environment, far/occluded objects generally reflect few laser signals back, thus generating less point cloud data. In other words, these objects are represented as relatively inconspicuous features in the corresponding feature map. As a result, they are usually miss-detected with low confidence scores less than the pre-defined detection threshold. As the $maxout$ function does not enhance the weak feature in fusion, these features will remain inconspicuous in the fused feature map, and keep those objects miss-detected. This issue needs to be addressed; otherwise, one major benefit of cooperative perception, namely extending the effective detection range of individual autonomous vehicles, will be compromised.

## III. CoFF: Cooperative Spatial Feature Fusion for 3D Object Detection

To address the above-mentioned limitations, we propose the Cooperative spatial Feature Fusion (CoFF) for cooperative 3D object detection on CAVs. CoFF effectively integrates feature maps so that the distinctive features are kept and enhanced, while noise features are suppressed. In essence, CoFF enables a vehicle (referred to as the receiver) to effectively utilize the supplementary information provided by another vehicle (referred to as the sender), and weighs the sender's feature map in the regions where its own feature map has a hard time detecting objects. With the increased weight from the sender's feature map, the noisy features on the receiver's feature map are eliminated by the *maxout* function, thus improving the object detection performance. Moreover, for objects that are either occluded or far from the receiver, CoFF enhances their corresponding features in the fused feature map, thus improving detection on these objects. The CoFF approach can be represented in the following equation:

$$\mathbf{F} = \{\mathbf{F}_3 \cup max\{\mathbf{F}_1, \mathbf{F}_2 \times X\}\} \times Y, \qquad (1)$$

where $\mathbf{F}_1$ and $\mathbf{F}_2$ are the overlapping areas of the two to-be-fused feature maps from the receiver and the sender, respectively, $\mathbf{F}_3$ is the non-overlapping area of the receiver's feature map, and $\mathbf{F}$ is the resulting fused feature map. $X$ is the assigned weight to the sender's feature map, $\mathbf{F}_2$, and $Y$ is the feature enhancement parameter.

### A. Information-based Spatial Feature Fusion

As the distance between a sender and a receiver increases, the overall object detection performance of F-Cooper decreases rapidly. This is due to the fact that F-Cooper treats all feature maps with the same weight regardless of their information contribution. For that reason, useless features contained in the receiver's feature map is also included in the fused feature map, which leads to a decrease in object detection performance. In this section, we propose the information-based spatial feature fusion to address this issue.

Information-based feature fusion consists of the following two steps: (1) semantic information measurement and (2) information-based fusion. When a receiver vehicle receives feature maps shared from a sender vehicle, it first measures the volume of new features that are contained in the feature maps by comparing them with the features in its own feature maps. Based on the measurement, the receiver applies a weight, $X$, to the received feature map before fusion. Therefore, the fusion counteracts the negative effect caused by the weak features in the receiver's feature map. To reduce this negative effect, we increase the weight of sender's feature maps in proportion to how much new semantic information it can contribute to the receiver on fusion results. The information-based fusion strategy can be expressed as

$$\mathbf{F}^i = max\left\{\mathbf{F}_1^i, \mathbf{F}_2^i \times X\right\}, \forall i = 1, 2, \cdots, 128, \qquad (2)$$

where $i$ denotes the index of a channel in a feature map, $\mathbf{F}_1$ represents the overlapping area of the feature map from the receiver, $\mathbf{F}_2$ represents the overlapping area of the feature map from the sender, and $\mathbf{F}^i$ is the $i$-th channel of the fused feature map.

Features within the overlapping area between the sender's and receiver's feature maps are similar to each other since they are generated from same physical region. Therefore, the similarity between the feature maps can be used to quantify the volume of new semantic information provided by the sender. The larger the similarity, the less supplementary information is provided by the received feature maps. We use L2 distance (also called Euclidean distance) between the corresponding features in the overlapping area of two feature maps to represent their similarity. A large L2 distance implies that the sender's feature map is able to provide a large volume of new semantic information. On the other hand, a small L2 distance suggests that the receiver's feature map is similar to the sender's, thus new information contributed by the sender would be limited. Besides similarity, the weight factor $X$ is also affected by the size of the overlapping area. A larger overlapping area results in a smaller weight. This is because a large overlapping area suggests a closer physical distance between the sender and the receiver, thus less new semantic information can be provided by the received feature map. The weight factor $X$ is calculated by the following formula

$$X = \begin{cases} S/(A_o/A) + 1.2 & S < 0.15, \\ S/(A_o/A) + 1.5 & 0.15 \leq S < 0.3, \\ 1.8 & S \geq 0.3 \end{cases} \tag{3}$$

where $S = \|\mathbf{F}_1^i - \mathbf{F}_2^i\|/(W \times H)$ and $\|\mathbf{F}_1^i - \mathbf{F}_2^i\|$ is the L2 distance between two feature maps which are represented by two vectors $\mathbf{F}_1^i$ and $\mathbf{F}_2^i$. $W$ and $H$ are the width and height of the overlapping area of two feature maps. $A_o$ is the size of the overlapping area, and $A$ is the size of entire feature map. The constant numbers in the above equations, e.g., $0.15$, $0.3$ and $1.2$, are derived from intensive experiments on our T&J dataset and our autonomous driving platform.

### B. Feature Enhancement

Objects with weak features in the fused feature map remain hard to detect with state-of-the-art 3D object detection models. Inspired by the recent work proposed in [13] where a binary classifier is used to predict the boundaries of objects, we discover that distant/occluded objects can be detected by increasing the difference between the values in a feature map that correspond to objects and the background. To this end, we propose the feature enhancement mechanism which can be represented as

$$\mathbf{F} = \{\mathbf{F}^i \times Y\}, \forall i = 1, 2, \cdots, 128, \tag{4}$$

where $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$ is the enhanced feature map, $\mathbf{F}^i$ represents the $i$-th channel of the 3-dimensional feature map produced by our information-based feature fusion mechanism. Here, the values of $W$, $H$ and $C$ denotes the width, height and the total number of channels of the fused feature map. With Eq. 4, the fused feature maps are enhanced by factor $Y$ before being passed to the RPN [14], which generates the classes and coordinates of detected objects. The enhancement increases the values in the feature map that represent objects, including distant and/or occluded ones. The values corresponding to the background in the feature map are mostly zero; as a result, Eq. 4 does not enhance background features.

The selection of the enhancement parameter, $Y$, is empirical, which depends on the quality of input 3D data. Analyzing the values magnitude of strong features on feature maps guides us in deciding the suitable values for the enhancement parameter $Y$. For the T&J dataset, which is collected by a 16-beam LiDAR, we find that an enhancement parameter $Y = 2$ or $Y = 3$ is adequate for enhancing weak feature of far/occluded objects, making the confidence score of most objects pass the detection threshold. For datasets with high quality of LiDAR data, e.g., KITTI [15], the selection of the enhancement parameter $Y$ can be decreased accordingly. The choice of $Y$ is also flexible for various road conditions. A larger $Y$ might be applied when real environments have more occlusions, such as heavy traffic scenarios, for a better detection performance. To avoid excessively enhance the fused feature map and the result of more false detection, we suggest the upper bound of choice $Y = 5$.

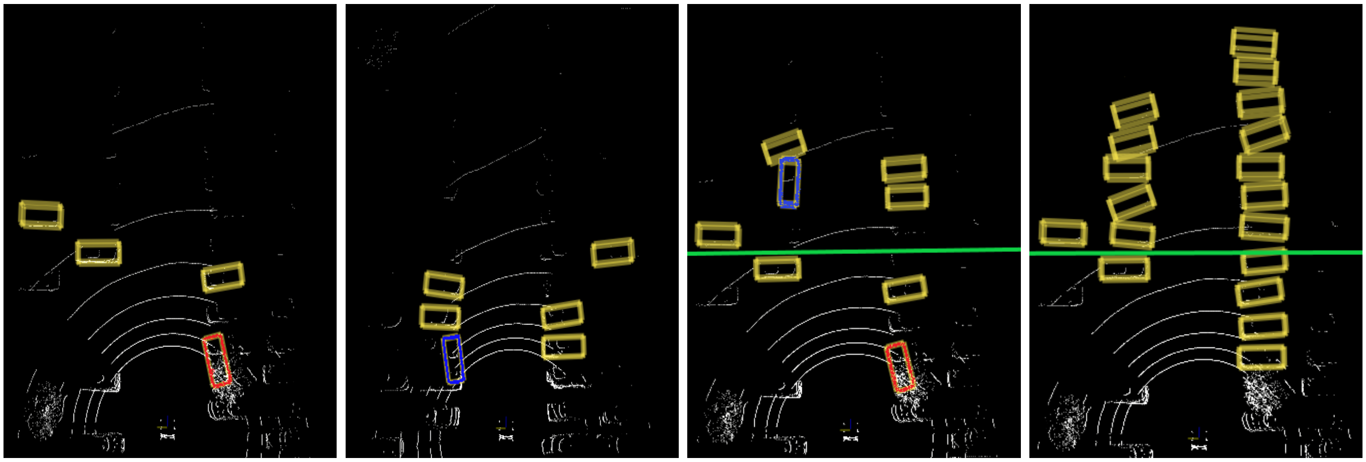
### C. Benefits of CoFF on Object Detection

In this subsection, we describe the benefits of applying CoFF in detecting 3D objects on autonomous vehicles.

*1) Extension of Detection Range:* One major benefit of applying CoFF to object detection is the extension of individual vehicle's detection range. As shown in Fig. 4(a) through (c), more objects in a larger area are detected after the sender's and receiver's feature maps are fused by F-Cooper. The detection range on the receiver is already extended by incorporating the feature map shared from the sender. However, the receiver struggles with detecting objects that reside at the boundary of the sender's sensing range, as exemplified by the miss-detected object shown in blue box in Fig. 4(c). Fig. 4(d) shows that after feature enhancement, CoFF is able to detect objects that are far away from both the sender and the receiver. As those features on the receiver's feature map are relatively weak, such improvement is mostly contributed by the enhanced features in the sender's feature map. The extension of the detection range is more significant in scenarios where less point cloud data is collected on an autonomous vehicle, e.g., due to occlusions or low-resolution data collected by a low-end LiDAR sensor.

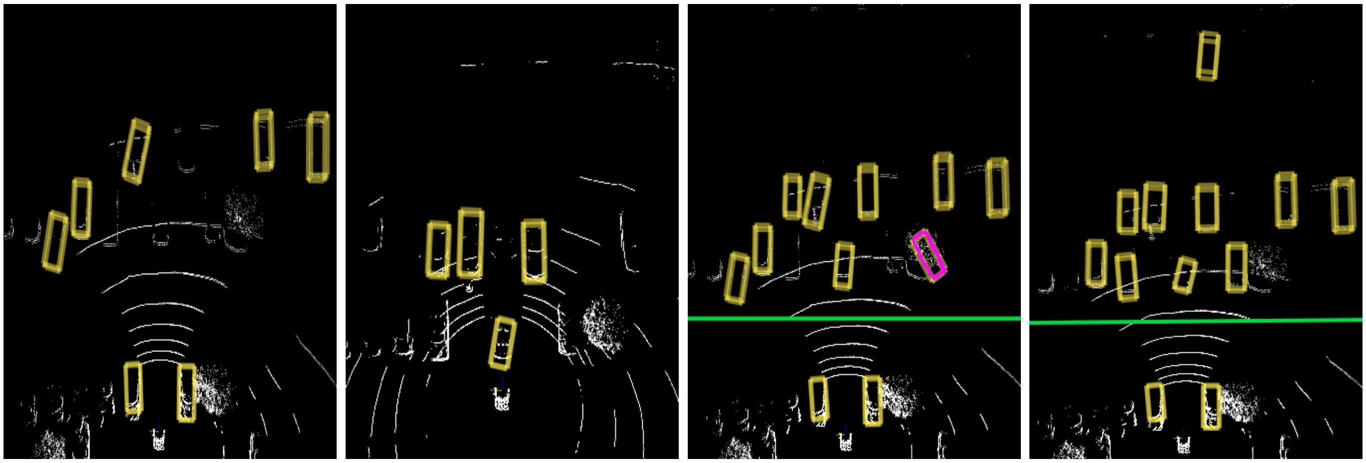
*2) Enhancement on Detection:* After enhancing the fused feature map, false detection originally caused by weak features can be reduced to a certain extent. A false detection example is shown in Fig. 4(b) where the blue box indicates that one vehicle is detected in this area. However, in reality there are two vehicles located in the blue box. Zooming into the blue box area, we identified the front portions of two vehicles that are parked parallel to each other. When the LiDAR sensor scanned this area, most of the laser signals are blocked by the vehicle(s) in front of them, resulting in little or no point cloud data collected from the rear portions of the two vehicles. Since the two vehicles are close to each other, the point cloud data collected from the front portions of the vehicles are treated as a whole, i.e., a single vehicle was mistakenly detected.

This issue can be addressed by CoFF so that the two vehicles can be separated, as shown on Fig. 4(d). This is because the weak features representing the rear portions of



(a) Vehicle 1 (Receiver) (b) Vehicle 2 (Sender) (c) F-Cooper detection result (d) CoFF detection result

Figure 4: False detection correction over individual detection results. Blue and red boxes show two false detections on F-Cooper detection results, where red box is not in the fusion area. Both can be corrected by CoFF. The green line splits the fusion (above green line) and non-fusion (below green line) areas in the fused feature map.



(a) Vehicle 1 (Receiver) (b) Vehicle 2 (Sender) (c) F-Cooper detection result (d) CoFF detection result

Figure 5: False detection correction over F-Cooper detection results. The magenta box is a false detection caused by F-Cooper's fusion function, which is corrected after the feature enhancement of CoFF. The green line splits the fusion (above green line) and non-fusion (below green line) areas in the fused feature map.

these two vehicles are enhanced by CoFF. With a stronger feature representing each vehicle, the RPN is able to detect two vehicles, instead of mistakenly treating the weak features as background. Moreover, the red box marked in Fig. 4(a) is another false detection case which originally is an overhanging tree and is not in the fusion region. Such false detection cases can also be reduced after feature enhancement, indicating that the enhancement is effective on individual detection models as well.

*3) Enhancement on Fusion:* Another type of false detection may arise when two weak features are improperly fused by the $maxout$ function. Because the $maxout$ function selects the most prominent features from the to-be-fused ones, some weak features may be enhanced in a wrong way. This is a unique problem to F-Cooper, because two feature maps are directly fused by the $maxout$ function. Fig. 5(c) shows an example of such false detection, in which a non-existing vehicle (depicted in the magenta box) appears after the feature maps generated by the sender and the receiver are fused by the $maxout$ function. Within the magenta box, there is a tree (the ground truth) which was falsely detected as a vehicle. With the proposed CoFF, the resulting feature map better represents the object in this area, as shown in Fig. 5(d), thus avoiding the false object detection.

## IV. Experiment and Results Evaluation

In order to compare performance with F-Cooper under the same settings, we carry out experiments with the T&J dataset used in [1] for cooperative 3D object detection. T&J is a real-world dataset collected by our autonomous vehicle for cooperative perception tasks on CAVs. All data are collected from a Polaris GEM e4 autonomous vehicle, equipped with a 16-beam LiDAR, six cameras, two radars, one IMU, and one GPS. Due to the limited cases available in the original T&J dataset, F-Cooper only provides performance evaluation in limited scenarios. During this work, we enrich the T&J dataset with the following three scenarios: road intersections, multi-lane roads and parking lots. Road intersection and multi-

| Scenario | Dataset | Cooper [2] | | F-Cooper [1] | | CoFF w/o Enhancement | | CoFF | | Improvement | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Near | Far | Near | Far | Near | Far | Near | Far | Near | Far |
| Multi-lane Roads | KITTI | 84.43 | 71.42 | 65.51 | 52.78 | 75.86 | 56.76 | 82.75 | 70.27 | 26.32 | 33.14 |
| Road Intersections | T&J | 80.21 | 66.37 | 54.28 | 28.57 | 62.75 | 34.61 | 72.54 | 59.25 | 33.64 | 107.38 |
| Parking Lots | T&J | 71.88 | 60.33 | 51.85 | 21.05 | 60.71 | 28.57 | 64.28 | 57.14 | 23.97 | 171.44 |

Table I: Precision comparison among Cooper, F-Cooper, CoFF, and the improvement of CoFF over F-Cooper on receiver (%).

lane road scenarios are used to simulate real driving conditions on road, and parking lot scenarios are used to evaluate CoFF in more cases with crowded environments. To evaluate CoFF on high-resolution 3D data, we select appropriate cases from the KITTI dataset, a well-known dataset for autonomous driving, to make a general comparison between CoFF and F-Cooper. In evaluation, we define objects within 20 meters from receiver AV to be in the "near" category, and objects beyond this range to be in the "far" category.

The evaluated dataset contains more than $1500$ and $200$ sets of data from the T&J and KITTI datasets, respectively. In experiment, we compare our CoFF solution with the feature fusion approach from F-Cooper, and the spatial feature map corresponds to a 3D space with a range of $[0, 70.4]$, $[-40, 40]$ and $[-3, 1]$ meters along the $x$, $y$, and $z$ axles. Our equipment to run CoFF is a desktop equipped with a NVIDIA Quadro P4000 GPU.

### A. Improvement from Fusion

CoFF consists of two parts, information-based fusion and feature enhancement. Information-based fusion is an improvement to the original *maxout* fusion used in F-Cooper, while feature enhancement is our novel approach to 3D object detection. We want to first quantify the improvement made possible from only information-based fusion. By doing so, we are able to find an upper-bound limit for the current 3D feature fusion strategy. Therefore, we discuss the improvements in detection precision without feature enhancement in this subsection.

Table I shows the detection precision of CoFF without feature enhancement, where both Intersection over Union (IoU) and confidence score threshold is set at $0.5$. For the "near" category, F-Cooper achieves a relatively high precision in open area scenarios such as multi-lane roads and road intersections. Due to the high similarity between the feature maps generated from two nearby vehicles, they are treated equally by *maxout*, i.e., the resulting fused feature map does not effect much on the detection results. Even so, our information-based fusion is able to obtain improvement of approximately $10\%$ for multi-lane roads and $8\%$ for road intersections. For parking lot scenarios, due to occlusion, both feature maps (from sender and receiver) contain less distinct features than that of open area scenarios. Thus, we see a decrease in detection precision for F-Cooper. For CoFF, the weights of the sender's feature map on fusion in these cases are much higher than those in an open area scenario, due to the low similarity between the two feature maps. Therefore, with more semantic information supplied by the received feature maps, CoFF is not affected much by occlusion, and still able to achieve approximately $9\%$ improvement on precision.

For the "far" category, as distance increases, the similarity between the two to-be-fused feature maps decreases accordingly. We see a lower detection precision with F-Cooper in all cases, especially in the parking lot scenarios. The improvements of our method over F-Cooper are slightly lower than those for the "near" category, which is about $7\%$ for cases from T&J dataset and $4\%$ for cases from KITTI dataset. Objects closer to the sender will have more distinct features on the sender's feature map. Such features also have a higher probability to be retained in information-based fusion. In the experiment, we found that most newly detected objects by the receiver is closer to the sender, which means our information-based fusion works well in retaining important features from the sender's feature map, thus improving cooperative perception. However, for objects that are far from both the sender and the receiver, their features are weak on both original feature maps and thus stay weak after fusion, therefore most of them remain undetected.

### B. Scenario Evaluation

To make a fair comparison of our CoFF method and how much improvement it yields, we re-implemented F-Cooper as our baseline. Moreover, we utilize Cooper fusion on the original 3D data as a reference for upper-bound limit in the evaluation. In the experiment, we follow the same approach as F-Cooper in designing our evaluation scenarios but with more evaluation cases. We report the precision by comparing the true detected vehicles against the ground truth, and set both the Intersection over Union (IoU) and confidence score threshold at 0.5 for detection. We also provide the corresponding improvement percentages over F-Cooper for a clear comparison.

In Table I, we first take a look at the baseline, which is F-Cooper. F-Cooper makes use of a non-weighted feature fusion which achieves a relatively high detection precision in the "near" category. We observe that F-Cooper is able to perform well on both multi-lane roads and road intersections, which shows that F-Cooper works well in open area scenarios and in short distance cases; the precision in these scenarios are over 54%. However, due to occlusions, the precision of F-Cooper for parking lots is not as good as compared to other scenarios. We see a precision drop in the parking lots cases with the precision being 51.85%. By comparison, the effect of occlusion is not as obvious on the CoFF method, CoFF is able to achieve a precision of 64.28% in the parking lot cases. For open area cases, the precision is above 72% on both multi-lane roads and road intersections, which shows a great improvement. The precision of CoFF in all scenarios is close to the upper-bound, Cooper, which means we could achieve a similar detection performance as raw sensor data fusion with much less data being transmitted between vehicles.

Moving to the "far" category, as distance increases, the precision of F-Cooper decreases rapidly. In the road intersection cases, the precision of F-Cooper is $28.57\%$. In the multi-lane road cases, since data from KITTI was collected by a high-end LiDAR, F-Cooper is able to achieve a precision of 52.78%. For more crowded scenarios such as the parking lot cases, the precision of F-Cooper drops to 21.05%. In contrast, the CoFF results clearly demonstrate the benefit of its feature enhancement improvement over F-Cooper. The precision of CoFF is above 57% for both scenarios from T&J dataset. Furthermore, CoFF does not present an obvious precision decline in parking lot cases. Benefiting from feature enhancement, most occluded objects are detected with a higher confidence score, suggesting that CoFF is not as sensitive to occlusion compared to F-Cooper.

In real-word deployment scenarios, a CAV participating in cooperative perception most likely would receive incoming sensor data from multiple, instead of one, close-by vehicles. For that reason, it is necessary to further evaluate the effectiveness of CoFF in multiple senders scenarios. For cases with two senders, CoFF preforms slight better in the "near" category, with approximately 18% higher precision improvement compared with the same cases with one sender. When more senders participate in cooperative perception, more semantic information can be obtained and fused from a larger physical perception region. We thus achieve a more prominent improvement on detecting "far" objects, with about 40% more improvement on precision. The improvement is greater in scenarios where there are more occlusions and CoFF is able to see a larger detection range in multiple senders cases. Due to space limitation, we here omit the detailed evaluation of CoFF for multiple senders cases.

### C. Precision Evaluation

Benefiting from the enriched T&J dataset, we are able to provide an overall accurate evaluation of CoFF's precision under various scenarios. Even though F-Cooper performs well on "near" objects, we are still able to see improvements with CoFF. **For CoFF, 70.51% of vehicles in the "near" category can be correctly detected, while F-Cooper achieves 52.13% within the same category.** For the "far" cases, CoFF has a even greater improvement over F-Cooper. **F-Cooper only has an average of 24.27% of detection precision in the "far" category, with most of the successful detection cases being within 35 meters. In contrast, CoFF is able to achieve a 58.18% detection precision, and the effective detection range reaches up to 50 meters.** This improvement is mainly from our feature enhancement approach. It is worth mentioning that the T&J dataset is collected by a 16-beam LiDAR, which generates fairly sparse point cloud data for far distance objects. Our experiment results suggest that CoFF's improvement works well in complicated road scenarios with limited points received by the LiDAR, and does not require dense 3D point cloud data.

To clearly illustrate how much improvements achievable with CoFF, we compare CoFF with F-Cooper in the "near" and "far" categories, respectively. Fig. 6 shows the Cumulative Distribution Function (CDF) of the precision improvement over the detection results of individual vehicles. As shown in the figure, for the "near" category, CoFF is able to achieve a 65% precision improvement for 80% of cases, while F-Cooper has about 32% of improvement when comparing with the detection results on individual vehicles. When it comes to the "far" category, CoFF is able to achieve a more distinct improvement. The improvement achieved by F-Cooper is around 20% for over 60% of cases, and is within 40% for almost 90% of cases. By comparison, CoFF is able to achieve about 120% of detection improvement over 80% of cases.

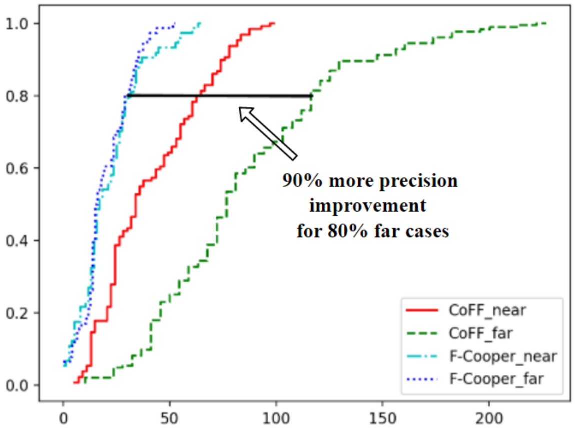


Figure 6: Cumulative Distribution Function vs. detection precision improvement percentages over individual vehicles

### D. Detection Range Evaluation

Through feature enhancement, CoFF is able to enhance weak features for far objects. Therefore, the effective detection range of CoFF is extended greatly over F-Cooper. We illustrate this improvement in Fig. 7 which shows drastic difference in detection range of the two approaches. For objects detected by F-Cooper, 83% of them are within 20 meters, which is in the "near" category. In contrast, for CoFF, the objects belonging to the "near" category only represent 61% of all detected. As we look deeper, the maximum detection range of F-Cooper is around 35 meters, while CoFF is up to 50 meters. For 80% cases evaluated by the two approaches, CoFF achieves an average of 11 meters of detection range improvement over F-cooper. As high resolution 3D data typically contains more dense points for far objects, this improvement is prominent in cases with low-resolution 3D data.

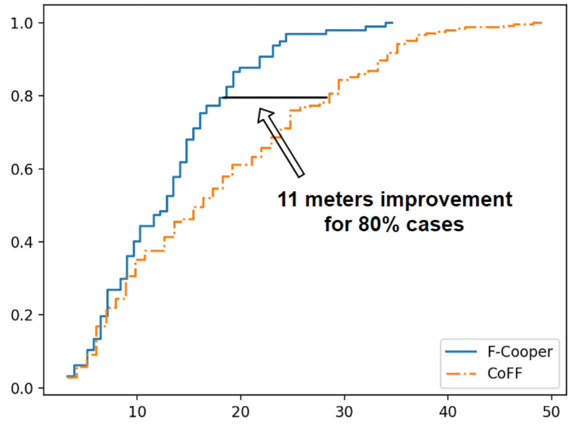


Figure 7: Cumulative Distribution Function vs. range of detected objects in meters

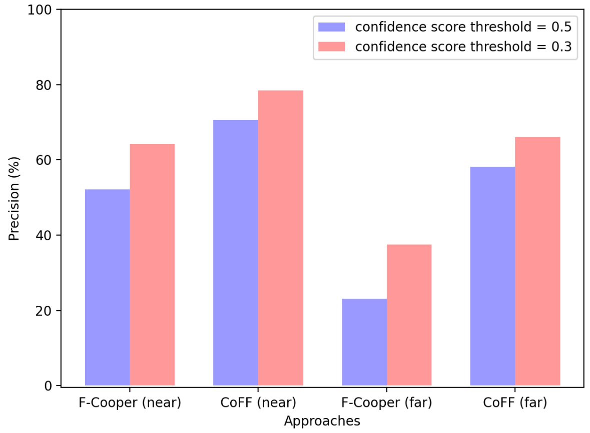


Figure 8: Precision comparison between F-Cooper and CoFF over different confidence score thresholds

### E. Detection Threshold Evaluation

Confidence score plays a great role in determining the making of a great detection model, as it directly impacts the model's final performance. Different thresholds impact the performance of certain models greatly, as seen in the performance of F-Cooper with different detection thresholds. However, CoFF is less sensitive to the confidence threshold than F-Cooper. Fig. 8 shows the precision comparison between F-Cooper and CoFF, with IoU $= 0.5$. We set the thresholds as $0.3$ and $0.5$, respectively. Thresholds over $0.5$ will cause a great drop in detecting far objects and thresholds below $0.3$ introduce a large number of false detections. As shown in Fig. 8, we see a great drop in precision when changing the threshold from $0.3$ to $0.5$ on F-Cooper, for both "near" and "far" categories. In contrast, there are minor changes on CoFF when the confidence score increases from $0.3$ to $0.5$, especially in the "near" category. We investigate the reasons in depth by analyzing the confidence scores of all detected objects by F-Cooper. As T&J dataset is collected by a low-end 16 beam LiDAR, the largest distribution interval of confidence score (of F-Cooper) resides between $0.3$ to $0.5$. With feature enhancement, most of them are converted to higher confidence scores greater than $0.5$. The features of most objects are enhanced well by feature enhancement, making our model less sensitive to the setting of detection threshold. In experiments, the confidence scores of most detected objects by CoFF are between $0.5$ and $0.7$. Since the improvement from a lower detection threshold is limited, we choose $0.5$ as the detection threshold of CoFF to avoid false detection.

## V. Conclusions

In this paper, we propose CoFF, a novel feature map based fusion approach for achieving cooperative 3D object detection on autonomous vehicles. CoFF consists of two parts: information-based fusion and feature enhancement. While the former allocates different weights on the received feature maps according to the amount of semantic information they contribute to the fusion, the latter enlarges the difference between the object and non-object areas on the feature map to achieve a better detection performance. Experimental results show that CoFF offers a better cooperative 3D object detection performance than F-Cooper while maintaining the same advantage of reduced data transmission, and does not require high-quality of 3D point cloud data.

## Acknowledgment

The work is supported by National Science Foundation grants CNS-1852134, OAC-2017564, ECCS-2010332, and the research grant from Fujitsu Network Communications.

## References

[1] Q. Chen, X. Ma, S. Tang, J. Guo, Q. Yang, and S. Fu, "F-cooper: feature based cooperative perception for autonomous vehicle edge computing system using 3d point clouds," in *Proceedings of the 4th ACM/IEEE Symposium on Edge Computing*, 2019, pp. 88–100.

[2] Q. Chen, S. Tang, Q. Yang, and S. Fu, "Cooper: Cooperative perception for connected autonomous vehicles based on 3d point clouds," in *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2019, pp. 514–524.

[3] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[4] S. Shi, X. Wang, and H. Li, "Pointrcnn: 3d object proposal generation and detection from point cloud," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 770–779.

[5] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander, "Joint 3d proposal generation and object detection from view aggregation," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1–8.

[6] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[7] M. Engelcke, D. Rao, D. Z. Wang, C. H. Tong, and I. Posner, "Vote3deep: Fast object detection in 3d point clouds using efficient convolutional neural networks," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 1355–1361.

[8] B. Li, "3d fully convolutional network for vehicle detection in point cloud," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 1513–1518.

[9] Y. Zhou and O. 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*, 2018, pp. 4490–4499.

[10] Y. Yan, Y. Mao, and B. Li, "Second: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, 2018.

[11] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 697–12 705.

[12] I. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, "Maxout networks," in *International conference on machine learning*, 2013, pp. 1319–1327.

[13] J. Wang, W. Zhang, Y. Cao, K. Chen, J. Pang, T. Gong, J. Shi, C. C. Loy, and D. Lin, "Side-aware boundary localization for more precise object detection," *arXiv preprint arXiv:1912.04260*, 2019.

[14] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

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

**Jingda Guo** is a Ph.D. candidate in the Department of Computer Science and Engineering at University of North Texas, Denton, TX, USA. He received his B.S. degree in Electrical Engineering from Northeast Electric Power University, China, and M.S. degree in Computer Engineering from University of Delaware, USA, in 2015 and 2017, respectively. His research interests include Internet of Things, Connected and Autonomous Vehicles, and Computer Vision.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3053184, IEEE Internet of Things Journal

**Dominic Carrillo** is a Ph.D. candidate in the Department of Computer Science and Engineering at University of North Texas, Denton, TX, USA. He received his B.S. degree in Computer Science and Mathematics from Sul Ross State University, Alpine, TX, USA. His research interests include Connected and Autonomous Vehicles, and Simultaneous Localization and Mapping (SLAM).

**Sihai Tang** is a Ph.D. candidate and a part of the Dependable Computing Systems Lab (DCS) in the Department of Computer Science and Engineering at University of North Texas, Denton, TX, USA. He received his B.S. degree in Computer Science from University of Texas at Austin in 2017, Austin, TX, USA. His research interests include Edge Computing, Federated Machine Learning, and File Systems.

**Qi Chen** is a Ph.D. candidate in the Department of Computer Science and Engineering at University of North Texas, Denton, TX, USA. He received B.S. and M.S. degrees in Electronic Engineering from Xidian University and Northwestern Polytechnical University, China, in 2006 and 2010, respectively. His research interests include Perception on Autonomous Vehicles, Edge AI and Internet of Things.

**Qing Yang** is an Assistant Professor in the Department of Computer Science and Engineering at University of North Texas, Denton, TX, USA. He received B.S. and M.S. degrees in Computer Science from Nankai University and Harbin Institute of Technology, China, in 2003 and 2005, respectively. He received his Ph.D. degree in Computer Science from Auburn University in 2011. His research interests include Internet of Things, Connected and Autonomous Vehicles, Network Security and Privacy.

**Song Fu** is an Associate Professor in the Department of Computer Science and Engineering at University of North Texas, Denton, TX, USA. He received his Ph.D. degree in Computer Engineering from Wayne State University in 2008. His research interests include Parallel and Distributed Systems, Cloud and Edge Computing, Connected and Autonomous Vehicles, System Reliability, and Machine Learning.

**Xi Wang** is a Research Scientist with Fujitsu Network Communications in Richardson, Texas USA. He received his B.S. degree in Electronic Engineering from Doshisha University, Japan in 1998, and received M.E. and Ph.D. degrees in information and communication engineering from the University of Tokyo, Japan in 2000 and 2003, respectively. His research interests include Software-Defined Networks (SDN), Packet Optical Networks, Vehicular Networks, Optical/Wireless Network Integration, Edge Computing/AI applications, and future Information and Communications Technology (ICT) fusion.

**Nanan Wang** is a Research Scientist at Fujitsu Network Communications in Richardson, Texas, USA. He received his B.S. degree in Computer Science and a M.E. degree in Computer Systems Organization from Jilin University, China. He also received his M.S. and Ph.D. degrees in Computer Science from The University of Texas at Dallas. His current research interests lie within Edge AI and Deep Learning for Computer Vision. His previous research experiences also includes Vehicular Networking, 5G Systems design, and Optical Network Systems. He has co-authored over 25 conference and journals papers, and holds 7 US patents.

**Paparao Palacharla** Paparao Palacharla is director of the Advanced Technology Lab at Fujitsu Network Communications. His current research interests include multilayer optical transport networks, SDN/NFV, 5G and Edge computing. He received his B.Tech. degree from the Indian Institute of Technology, Kharagpur, and his Ph.D. Degree from the University of Iowa in electrical and computer engineering. He has previously worked at the National Research Council, Nortel Networks, and Fujitsu Network Communications in the areas of optical signal processing, optical interconnects, and optical networking. Dr. Palacharla is a senior member of the IEEE and has served as a technical program committee member for OFC/NFOEC,GLOBECOM, and other conferences.