electronics

MDPI

Article

Objects Detection Using Sensors Data Fusion in Autonomous Driving Scenarios

Razvan Bocu [1,*,†,‡], Dorin Bocu [1,‡] and Maksim Iavich [2,‡]

[1] Department of Mathematics and Computer Science, Transilvania University of Brasov, 500036 Brașov, Romania; dorin@bocu.ro
[2] Department of Computer Science, Caucasus University, Tbilisi 0102, Georgia; miavich@cu.edu.ge
* Correspondence: razvan.bocu@unitbv.ro; Tel.: +40-732011010
† Blvd. Iuliu Maniu Nr. 50, Brasov, Romania.
‡ These authors contributed equally to this work.

Abstract: The relatively complex task of detecting 3D objects is essential in the realm of autonomous driving. The related algorithmic processes generally produce an output that consists of a series of 3D bounding boxes that are placed around specific objects of interest. The related scientific literature usually suggests that the data that are generated by different sensors or data acquisition devices are combined in order to work around inherent limitations that are determined by the consideration of singular devices. Nevertheless, there are practical issues that cannot be addressed reliably and efficiently through this strategy, such as the limited field-of-view, and the low-point density of acquired data. This paper reports a contribution that analyzes the possibility of efficiently and effectively using 3D object detection in a cooperative fashion. The evaluation of the described approach is performed through the consideration of driving data that is collected through a partnership with several car manufacturers. Considering their real-world relevance, two driving contexts are analyzed: a roundabout, and a T-junction. The evaluation shows that cooperative perception is able to isolate more than 90% of the 3D entities, as compared to approximately 25% in the case when singular sensing devices are used. The experimental setup that generated the data that this paper describes, and the related 3D object detection system, are currently actively used by the respective car manufacturers' research groups in order to fine tune and improve their autonomous cars' driving modules.

Keywords: objects detection; autonomous driving; sensors data

Citation: Bocu, R.; Bocu, D.; Iavich, M. Objects Detection Using Sensors Data Fusion in Autonomous Driving Scenarios. *Electronics* **2021**, *10*, 2903. https://doi.org/10.3390/electronics10232903

Academic Editors: Ohyun Jo, Byungchang Chung and Minhoe Kim

Received: 18 October 2021
Accepted: 22 November 2021
Published: 24 November 2021

1. Introduction

The role of autonomous driving components is to collect relevant data that creates an accurate overview concerning the driving environment. Consequently, the precise operation of these components is essential for the reliable autonomous driving of relevant vehicles in environments with various degrees of complexity. The improper detection of environmental components may result in tragic incidents. These are caused by failure of the autonomous driving components to properly detect and classify dangerous objects and patterns on the cars' path.

The detection of 3D objects is placed at the core of the autonomous driving components. This process involves the estimation of 3D bounding boxes that define the objects' spatial position and orientation, and identifying the category to which the objects belong in the environment. The detection of 3D objects is usually performed using machine learning techniques, and the assessment may consider either real-world or synthetic data. As an example, there are well known datasets, such as KITTI [1], which include images gathered by frontal cameras, and also data that define the 3D boxes annotations. It can be stated that rich texture images, which are generated by the cameras, are essential for the proper and efficient classification of 3D objects. Nevertheless, light detection and ranging laser-based cameras (Lidar), along with depth cameras, may provide useful data that can be used in

order to determine the objects' spatial position and orientation. In this context, it can be asserted that many detection schemes consider data that is gathered from multiple sensors, in order to increase the actual detection performance and accuracy.

Nevertheless, the consolidated sensor data may prove vulnerable to certain functional problems. This category of problems includes occlusion, restricted perception horizon that is related to a limited field of view, and also the low-point density relative to distant regions. Thus, a solution to these kinds of problems may be represented by cooperative data collection that includes various sensors. Existing reported contributions demonstrate achievements, although relatively limited, concerning lane selection, maneuver coordination, and automated intersection crossing. This paper considers the goal to detect 3D objects, which pertains to the data that is cooperatively acquired from multiple sensors. Thus, instead of using the disparate data that is collected from individual sensors, the solution is to combine the collected data (cooperative perception). This determines significant advantages, such as an increased perception of the horizon, and a reduced occurrence of noisy data.

The 3D object detection, which is based on cooperative perception, may be accomplished through two particular modalities, late fusion (LF) and early fusion (EF). The modalities are named in regard to temporal sequence, as the fusion may occur before or after the 3D object detection stage. In late fusion, each data acquisition is regarded and processed in an independent manner. The detected 3D boxes are combined (fused) into a resulting 3D entity. Moreover, early fusion processes the acquired raw data and combines (fuses) it prior to the detection stage. Consequently, late fusion schemes are classified as high level detection schemes, while early fusion belongs to the category of low level detection schemes. While both approaches are essentially capable to enhance the perception horizon and field-of-view, only the early detection scheme is able to efficiently use the raw image data. Thus, let us consider the typical example of a vehicle, which is partially covered (occluded) as it is perceived from two different detection positions. In such a scenario, each particular sensor is determined by a certain occlusion model, which produces an essentially unreliable or even unsuccessful detection. Consequently, the combined data may provide the possibility to successfully detect the 3D object.

Considering the aspects presented above, this paper proposes the detection of 3D objects through both schemes, the late and early detection schemes. This essentially results in a detection scheme that supports the late and early detection models, which can be used as required at the level of the central data processing components. The contribution that is presented in this paper considers a system, which aims to produce a precise construction of roads, regardless of their complexity. This includes problematic road components, such as roundabouts and T-junctions. The system was tested using the data obtained from experimental sensors that are installed at road level. This approach is capable of providing the necessary data to allow safe autonomous driving scenarios to occur in virtually any scenario. Consequently, safe autonomous driving can be conducted on even the most complex road segments. This paper reports on the following contributions.

- Two cooperative schemes for the detection of 3D objects are described, which consider the mentioned fused detection schemes, and a custom training model.
- The algorithmic routines are optimized through a comprehensive experimental process. This allows for the implemented software components to perform better than existing similar approaches.
- The dataset that is used for the assessment procedure is obtained through cooperation with automotive industry partners.
- The late and early fusion schemes are comprehensively assessed using real-world data.
- Additionally, the importance of the data acquisition sensors' deployment is analyzed and discussed, with significant results concerning the deployment of such a system in a real-world scenario.

The rest of this paper is structured considering the following parts. The following subsection briefly describes the structure of the investigative process. This has the role

to facilitate understanding of the detailed scientific presentation that is contained in the following sections. Thus, in Section 2, the most relevant existing contributions are presented, with a comparative assessment relative to this paper's reported contribution. It also includes a review of the most relevant fusion based detection schemes, which have been studied during the initial stages of the research. The following section describes the cooperative 3D objects detection model, together with an explanation of the relevant fusion schemes. Following this, the considered real-world dataset is described, and the most relevant features of the training process are presented. The outcomes of the practical evaluation process are also described, with an emphasis on some essential theoretical and practical aspects. The last section concludes the paper.

Structure of the Scientific Contribution

The automotive industry partners that supported the creation of the field experimental setup expressed their concern regarding the suboptimal behaviour of 3D objects detection systems, which use the standard fusion based detection schemes: Multi-View 3D Networks (MV3D) [2], Aggregate View Object Detection (AVOD) [3], and Frustum PointNet (F-PointNet) [4]. Thus, the issues that were reported pertain to the unsatisfactory level of detection accuracy, and to the slow detection of the 3D objects, which were preventing some real time autonomos driving decisions being made. This can produce a wide array of autonomous driving issues, which range from mild delays adapting to traffic sign indications, to catastrophic collisions with other vehicles. Therefore, the investigative effort started with the assessment of the basic fusion based models, which are MV3D, AVOD and F-PointNet.

Following this, the integrated 3D objects detection system was designed. This is intended to collect the data using road side data acquisition sensors, along with the cars' front view sensors, when they are available. The road side data acquisition sensors possess lasers-based Lidar capabilities, but less expensive plain optical sensors can be mounted. Furthermore, the software engineering development process involved the design and implementation of the central data processing components, which realize the actual combination and processing of the data that are collected by the individual sensors. Furthermore, the received data are processed, and the 3D objects are identified through the determination of the related bounding boxes. Finally, the central data processing components send the results of this data processing to the autonomous vehicles that drive in the neighbourhood. This overall process occurs in a real-time fashion, as is demonstrated by the outcomes of the experimental assessment that is presented in this paper. Thus, none of the delays that we have initially experienced during the assessment of the three reference fusion based schemes occurred. Consequently, the proposed integrated 3D objects detection system addresses the fundamental problems that were raised by our automotive industry partners. Therefore, this research prototype and the developments that are reported in this paper are actively considered by the relevant research teams that work on the development of autonomous driving software routines.

2. Relevant Existing Contributions

This section includes relevant information concerning the most important similar existing contributions. Thus, the first subsection discusses the relevant detection models, which mostly pertain to systems that are based on single-sensor approaches. The next subsection presents the most relevant detection schemes for 3D objects. Consequently, a comparative analysis with the work that is described in this paper is conducted.

2.1. Remarks Concerning 3D Objects Detection Models

The relevant models may be placed into several categories considering the input data modality as criterion. Thus, there may be colour images that are obtained from monocular cameras, point clouds, which are generated by laser-based Lidar devices or depth cameras, and finally a combination of both these approaches. It is relevant to note that monocular

cameras do not generally produce sufficient depth cues [5], unless a moving camera is used. Consequently, monocular cameras will be essentially disregarded as an acceptable solution in this concise review, as they cannot provide suitable data for the detection of 3D objects. Consequently, the point clouds may be considered in order to estimate the spatial placement and orientation of the objects relative to sensibly higher accuracy levels than the approaches that consider data provided by monocular cameras [6]. The interested reader may consult a dedicated and certainly more comprehensive review on this problem in [6,7].

The 3D objects detection models that consider point clouds generally process the related point cloud in the form of a Bird-Eye-View [8], or even through so-called cylindrical coordinates [9]. These have the role of generating a proper 3D structure that can be processed by convolutional neural networks, which realize the effective object detection. Thus, the representation of the points into a fixed size input tensor determines the generation of the definitive 3D bounding boxes during one forward pass through the convolutional neural network [10]. Nevertheless, this technique may produce the loss of some useful information [11] due to sub-optimal representation choices [12]. As compared to the mentioned representation techniques, Voxelnet [13,14] uses the raw 3D points in order to generate a properly structured input. Additionally, PointRCNN [15] and STD [16] represent stage detectors, which use PointNet [17,18] with the purpose to determine better points configurations.

The contribution that is reported in this paper is centred on the problem of designing a 3D object detection model with cooperative capabilities. Considering the computational performance of Voxelnet [14], useful design suggestions [19,20] are considered. Furthermore, considering the experiments that we conducted, we are able to state that this approach allows us to reduce the transmission bandwidth between the data collection sensors and the central data processing components. This is an important advantage [21,22] considering the large amount of data that is being produced during a real-world autonomous driving session [23].

2.2. Essentials Concerning 3D Objects Detection Schemes

The study that is reported in [24] describes a neural network architecture, which may be useful for the detection of 3D objects in point clouds that are sparse. Furthermore, this study discusses communication costs, robustness and error handling, and it demonstrates that 3D objects can be detected more efficiently through cooperative detection strategies. Furthermore, the research that is reported in [25] discusses feature-level fusion schemes and assesses the trade off that can be established between processing time, bandwidth usage and detection performance. Additionally, the authors of paper [26] research the problem of trust, which is necessary in order to prevent attacks on 3D object detection systems [27] that use cooperative detection models. The proposed 3D object detection model is assessed using a synthetic dataset that is generated using a game engine that renders urban scenes.

The contribution that is reported in this paper differentiates in several respects. For example, the studies in [2,28] propose a peer to peer data transmission protocol, and the data processing occurs locally. We describe a central data processing approach, which combines the data that are generated by the data acquisition sensors [3,29]. Furthermore, the data are processed in a shared fashion. Moreover, this paper studies two relatively complex road scenarios, in connection to any autonomous driving process, the roundabout and the T-junction. Additionally, this paper assesses the impact that the number of data acquisition sensors and their placement has on the performance of the entire 3D object detection system. This is often a theoretical and practical problem that is neglected by similar studies.

2.3. Taxonomy of 3D Objects Detection Methods

The relevant scientific contributions suggest that the 3D object detection methods are classified into three categories [4,30]. There are monocular image, point cloud, and fusion based methods. The information that is presented in Table 1, which comparatively describes the available 3D object detection methods, suggest that the fusion schemes represent the optimal choice, both considering the accuracy of the detection, and the necessary computational resources that are required [31]. Consequently, the following subsection discusses relevant data fusion schemes, which have been initially considered during the initial stages of the research process that is reported in this paper.

Table 1. Taxonomy of 3D Objects Detection Methods.

Type	Advantages	Disadvantages	Opportunities
Monocular	Single RGB images to predict 3D object bounding boxes. Finds 2D bounding boxes on the image plane then extrapolate them to 3D.	Lacks depth information	Convolutional Neural Networks may be used to increase detection accuracy.
Point-cloud Projection	Projects point clouds into a 2D image and uses existing models for object detection on 2D images to generate 3D bounding boxes.	Information loss, prevents explicit encoding of spatial information.	Machine learning models may be used to improve detection accuracy
Point-cloud Volumetric	Considers 3D point clouds and fully convolutional neural networks.	Extensive 3D convolutions, computationally inefficient.	Region proposals may be used to improve accuracy.
Point-cloud PointNet	Generates predictions of 3D bounding boxes based on feed-forward networks.	Considers whole point cloud, difficult to establish region proposals.	Methods based uniquely on point-cloud should be assessed.
Fusion	Combines both front view images and point clouds to specify a robust detection.	Requires calibration of sensors.	State-of-the-art, different environment conditions should be investigated.

2.4. Literature Review of Existing Fusion Based Methods

It has already been stated that point clouds do not offer texture information, which is essential in order to determine the class of a certain object during the detection process. Furthermore, monocular images are not able to determine depth values, which are essential in order to precisely determine the 3D objects' spatial location and physical size. Moreover, the density of the point clouds reduces directly proportional to the distance increase from

the sensor. Thus, the fusion schemes represent the solution, which aims to address the enumerated shortcomings. Let us recall that there are three categories of fusion schemes [2].

- Early fusion (EF) schemes—They determine a detection process, which involves the combination of modalities during the inception stage.
- Late fusion (LF) schemes—This approach determines a detection process where modalities are processed separately until the last stage, which determines the fusion.
- Deep fusion (DF) schemes—This type of process combines the modalities in a hierarchical fashion relative to the layers of the respective neural network.

The authors of paper [28] assess the fusion process considering all the phases of a 3D pedestrian detection process. Thus, the described model is based on two inputs: a depth frame, and a monocular image object. Their research asserts that the late fusion scheme determines the optimal performance, while the early fusion may be used if the implied performance penalty is not particularly significant in the context of the particular use case.

The scientific literature describes the possibility of using the point cloud projection method, which is based on the utilization of front facing cameras that feature additional RGB channels. Thus, the papers [2,3] consider region proposal networks in order to generate the required 3D regions of interest (RoI). The method that is called MV3D [2] considers bird-eye and front view projections of lasers-based Lidar points, and a front view camera that features the appropriate RGB channels. The network is composed of three input branches, which correspond to each view, relative to the VGG-related feature extractors (Visual Geometry Group) [29]. It is relevant to note that the 3D proposals, which are determined relative to the bird-eye view features only, are projected to the feature maps that pertain to each particular view. The involved features are combined considering a deep fusion scheme. The model outputs the result of the classification process, and the vertices that define the 3D bounding box. The authors state that the deep fusion approach generates the best performance in the context of the tests they conducted, while it also allows for the features' aggregation to occur in a more flexible manner.

We have reviewed another method, which is called AVOD [3]. This represents the first approach that considers an early fusion approach, which is featured by a merge between the bird-eye view and the respective RGB channels. The representation of the input data is similar to the one that is described as MV3D in paper [2], with the remark that only the image data input branches, and the bird-eye view are considered. Nevertheless, the performance assessment that we conducted demonstrates that data processing through the convolutional network determines a certain loss of the processed image objects' details, which prevents the proper detection of small objects. The authors attempt to improve the accuracy of their detection model by using the Feature Pyramid Networks [31], which determine an upsampling process. The authors state, and we have partly confirmed, that this approach is efficient in the case of poorly illuminated or snowy scences.

There is another strategy that is based on the utilization of monocular images in order to obtain the required 2D image candidates, which are consequently transposed to the 3D coordinate system. The approach that is called Frustum PointNet [4] suggests the expected region proposals, and it is also able to perform the required classification and the determination of the bounding boxes. Thus, the 2D boxes are transposed to the 3D space using the camera calibration parameters, which determine the frustums region proposals. Each frustum encompasses a set of points, which are selected and segmented through the utilization of a PointNet instance, which effectively removes useless background data. Following this, the effective classification is conducted by another PointNet instance. Furthermore, another approach is described in paper [30], which suggests the selection of the points that are placed inside the related detection box. Then, the chosen points are used in order to perform the model fitting, which results in an initial 3D proposal. This proposal is computed considering a convolutional neural network with two stages, which refines the existing data, and generates the final 3D box, together with the confidence score. These approaches behave sub-optimally in certain environmental conditions, such as poor lighting, due to the usage of monocular images.

The survey that we conducted demonstrates that the selection of data fusion schemes for the design and implementation of the 3D objects detection routines that are described in this paper is completely justified. The following section further discusses relevant aspects and contributions that pertain to the data fusion schemes.

2.5. Relevance of Data Fusion Schemes

The work that is presented in [19] discusses classification regarding various data fusion approaches. The most prevalent techniques are placed into three main categories: data association, state estimation, and decision function. In the context of this classification, the fusion schemes may be classified as complementary, if the data acquisition sensors determine exclusive field-of-views. Furthermore, the fusion schemes are regarded as redundant if sensors with overlapping field-of-views are used.

The survey contribution that is presented in paper [19] discusses three classes of point cloud fusion approaches, which may be used for remote data collection. Point cloud level is first discussed, which features several points or features that are added to the initial point cloud set. The voxel level is also reviewed, which features point clouds that are combined (fused) through a voxel representation. Additionally, the survey also discusses the features level, which manifests relative to the detected 3D objects.

It is important to state that the approach that is presented in this paper belongs, in part, to the first category. It is also significant to mention that it proposes unique features that are presented in the next sections.

3. Presentation of the Proposed Model

The proposed model, which represents an extension and algorithmic optimization of the Frustum PointNet (F-PointNet) fusion based method, considers n data acquisition sensors. Each of them is capable of providing depth sensing capabilities, which include laser-based Lidar, and also depth assessment functions. Furthermore, each data acquisition component features a local data processing unit (processor). Let us recall that we proposed a centralized data acquisition model for the detection of 3D objects. This implies that the dispersed data acquisition sensors send the collected data to the central processing components using wired or wireless data links. Moreover, it can be stated that the sensors that were used in order to collect the data are properly calibrated. This means that their precise spatial position and orientation are communicated to the central data processing components. The original F-PointNet fusion based model usually processes the front view images. This constraint implies that this approach has difficulties in properly detecting 3D objects in certain regular environmental conditions, for example at night time. Although F-PointNet is regarded as one of the most reliable fusion based methods, this drawback, that manifests often in practice, is unacceptable for our automotive industry partners. Therefore, we have extended and improved the original F-PointNet fusion based method, in such a way that the model that we propose is able to accurately detect 3D objects in virtually any environmental conditions. The following sections describe and assess the proposed model.

The central data processing components possess enough computational power in order to process and combine the data that is received from the data acquisition sensors. Additionally, the central data processing components are able to synchronize and send useful data to nearby vehicles using a radio system. This implies that the vehicles that are used possess the necessary hardware components in order to receive the radio signals from the central data processing components. Furthermore, the enrolled vehicles possess the necessary computational capabilities in order to locally process relevant data, which pertains to the implied autonomous driving processes. This implies that data, which relate to environmental perception and trajectory control, are processed. It is relevant to note that the central data processing components are not responsible for the transmission of actual trajectory control signals for the enrolled vehicles. Thus, the respective vehicles use the data that are supplied to them by the central data processing components in order

to perform the necessary data processing and make their own decisions concerning the proper control of their trajectories. The rationale behind the presence of the central data processing components, as an architectural component of the proposed approach, is to offload computationally expensive data acquisition and processing operations from the actual vehicles, and allow them to make more precise decisions.

3.1. The Data Curation Phase

The 3D objects detection model that is described in this paper processes point cloud data, which can be provided by laser-based Lidar devices or depth cameras. The laser-based Lidar devices are able to generate point clouds natively. The images that are generated by depth cameras should go through a post-processing phase in order to generate the required point clouds. It is necessary to note that each sensor, which is part of a certain hardware assembly, determines data points that pertain to its own coordinates system. Consequently, the respective acquired data must be translated to the global coordinates system, which can be properly handled by the central data processing components. This translation is based on the operation of rotation, which is followed by a proper translation phase. This effectively associates data points that are generated by the individual data acquisition sensors to the central coordinates system, which is determined by the inverse of the extrinsic matrix [20] that is associated to each data acquisition sensor. Let us consider the coordinates (a, b, c) of a 3D point in the coordinates system of sensor s. Consequently, the global reference point (a_r, b_r, c_r) may be determined considering the following formula.

$$\begin{pmatrix} a_r \\ b_r \\ c_r \\ 1 \end{pmatrix} = Mat_i^{-1} \begin{pmatrix} a \\ b \\ c \\ 1 \end{pmatrix} = [Rot_i | Trans_i] \begin{pmatrix} a \\ b \\ c \\ 1 \end{pmatrix}$$

Here, Mat_i represents the extrinsic matrix of sensor s. This can be decomposed into a rotation matrix Rot_i, and also into a translation vector $Trans_i$. The extrinsic matrix Mat of a sensor s, and consequently Rot and $Trans$ are adequately determined through a calibration process. This can prove to be a challenging process in practice, considering that Mat depends on the spatial position and orientation of the implied data acquisition sensors. Thus, the accuracy of the obtained result directly depends on the measured values of these variables. Therefore, considering that the sensors are placed on mobile data acquisition nodes, which are typically represented by a vehicle, any error that pertains to the localisation data of the mobile node determines alignment issues concerning the fused point cloud. This may imply that non-existing 3D objects are detected, or existing 3D objects are not detected. Let us recall that the model that is described in this paper considers data that is collected by sensors that are placed at the side of the road segment, which is used in order to conduct the necessary experiments. The interested reader may obtain more information about this interesting topic by reading the papers [21,22]. The successful transformation of the point data that are collected by the individual data acquisition sensors into the coordinates system that is considered by the central data processing system, implies that any data which do not specifically pertain to the detected 3D object and its vicinity are removed.

3.2. Discussion Concerning the Early Fusion Scheme

The basic architectural and logical structure of this scheme is illustrated in Figure 1. It can be immediately noticed that it is based on the fusion of the point clouds, which are generated by n data acquisition sensors. This allows for the data that pertains to the 3D objects in the analyzed environment to go through an aggregation process. The experiments that we conducted show that this significantly increases the successful detection of the 3D objects that are affected by noise. The data processing workflow goes through a series of distinct phases. Thus, the preprocessing phase is the responsibility of each individual data acquisition sensor. This phase generates a cloud of n points that is referred

in the coordinates system that is understood by the central data processing components. Furthermore, each individual point is transmitted to the central data processing system, which aggregates the individual data points it receives into a single data point, and which is then sent as input to the 3D object detection component. This particular component produces, as an output, a list of objects, which defines the required 3D bounding boxes. Consequently, these data are sent over to the enrolled autonomous vehicles.

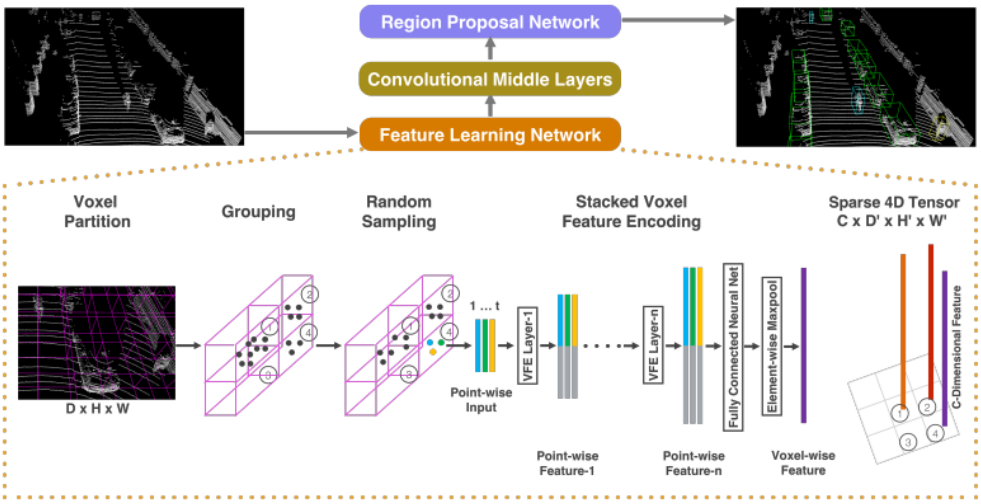

Figure 1. The early fusion scheme.

3.3. Description of the Objects Detection Component

This component is composed, from an architectural perspective, of three sub-components. Thus, it includes a feature learning network, multiple convolutional middle layers, and a Localized Proposal Network (LPN).

The feature learning network has the role of transforming the 3D point cloud data into a constant size representation, which can be efficiently processed by the convolutional layers of the neural network. Let us note that the original Voxelnet model considers the lasers-based reflection intensity channel, as well as the 3D spatial coordinates of the detected point, let us refer to them as (a, b, c). The 3D objects detection model that is described in this paper exclusively considers the spatial coordinates. This enables the data processing phase to efficiently produce the points that can be easily understood by the central data processing components. Thus, the input point cloud is separated into voxels of equal size, let us refer to them as (vox_a, vox_b, vox_c). Here, the components represent the width, length and height, respectively. Considering each voxel, a set of t points is selected in order to create a proper feature vector. Additionally, let us consider that T is the threshold concerning the maximum number of points for each individual voxel. If t is greater than T, a randomized sample of T points is generated. This has the role of decreasing the computational load, and improving the balance regarding the points distribution between different voxels. Moreover, the coordinates of these points are transmitted to a chain of Voxel Feature Encoding (VFE) layers. Thus, each particular VFE layer is composed of fully connected layers, local aggregations, and also max-pooling operations [14]. The output of this neural network is represented by a 4D tensor, which is indexed considering the following features of the voxel: dimension, height, length and width.

The middle layers in the convolutional neural network are particularly important. In essence, they add to the data processing pipeline three additional stages relative to the 4D voxel tensor, which has been presented. These stages include spatial data from the voxels in the vicinity, which add the mandatory three-dimensional context to the considered features map.

The Localized Proposal Network (LPN) receives the tensor that is obtained. The LPN network is structured considering three stages of convolutional layers in the neural

network. They are immediately processed through three supplementary transposed convolutional layers. Consequently, a high resolution features map is obtained. The generated features map is considered in order to produce two output branches. The algorithmic model considers a confidence score, which indicates the probability for a 3D object to be present in a certain analyzed scene. The algorithmic model also uses a regression map, which determines the position, orientation, and size of the processed bounding box. The interested reader may find out additional interesting information considering alternate similar approaches in [23].

3.4. Discussion Concerning the Late Fusion Scheme

The late fusion scheme combines the list of 3D bounding boxes, which is generated by each data collection sensor. It is possible that an object is not properly detected as a consequence of the noise levels in the acquired image samples, or because of the occlusion. Consequently, the central data processing components become unable to properly process the collected data. The workflow starts with the preprocessing of each individual point cloud. The output is provided to each data collection sensor, which generates a list of 3D objects that are defined by their 3D bounding boxes. Consequently, the entire set of detected 3D objects is sent to the central data processing components, which combine them into a unitary set of objects. We have determined that the unitary set of objects may contain multiple representations of the same object entities, as they are detected by several data acquisition sensors. Therefore, we have developed an algorithm that determines the overlap between different bounding boxes. In the case when the overlap between any pair of bounding boxes is above a specific threshold, the bounding box that is featured by the lower value of the confidence score is removed from the set. Thus, the value of this score essentially represents the probability with which an object is present inside a particular bounding box. The experiments that we conducted demonstrate that this approach successfully eliminates the bounding boxes that overlap. Additionally, the supplementary processing that it requires does not induce a substantial overhead on the overall 3D objects detection process. It is essential to note that the central data processing components conduct the required data processing operations. Consequently, the list of detected 3D objects is transmitted to the autonomous vehicles that move in the neighbourhood, which use this data in order to make their autonomous driving decisions.

4. The Experimental Dataset

The initial stages of the research that is reported in this paper suggested that there is no proper dataset that may be used in order to experimentally assess and tweak the model that is presented in this paper. Therefore, we asked and obtained from our industrial partners the permission to use automotive data, which has been generated during the testing of their respective vehicle prototypes. It is relevant to note that the contributions that are reported in [16,17] simulate a testing environment, which implies that two vehicles share the data that is produced by each of them through the consideration of point clouds that are part of the dataset KITTI [1]. Nevertheless, these approaches are flawed by a reduced number of driving scenes, where the 3D objects are mostly perceived from a static perspective. Therefore, this dataset is not precisely suitable in order to fully assess the detection of the 3D objects in a fully featured autonomous driving scenario. Consequently, the experimental dataset that is considered in this paper addresses these shortcomings, and it also has the advantage that it includes real-world data, as opposed to most of the existing similar contributions, which consider synthetic data that is generated with software tools like CARLA [25], or through the utilization of other approaches.

The experimental dataset contains data that were gathered by data acquisition sensors, which are statically placed at the edge of the road. The data acquisition sensors' placement is adequate in order to create the two basic scenarios, the roundabout and the T-junction. The data acquisition sensors are capable to capture depth and RGB image data at a resolution of 640 $\times$ 480 pixels, while the horizontal field of view is provided at a 90 degree angle. The

area of the T-junction is monitored by ten data acquisition sensors, which are mounted on vertical masts with a height of 4.3 m. The balanced acquisition of data is ensured by the setup of the sensors, which implies that five of them target the incoming direction, while the other five point to the T-junction's opposite direction. Furthermore, the roundabout scenario considers twelve data acquisition sensors, which are placed on masts at a height of 6.1 m. Considering the roundabout data acquisition sensors, it can be stated that six of them scan the incoming road lanes, while the other six sensors monitor the outgoing road lanes. The placement of the data acquisition sensors, considering both scenarios, was optimized through an empirical onsite process.

The experimental dataset is composed of four independent sections. Thus, two data collections pertain to the T-junction scenario, while the other two are related to the roundabout scenario. Each data collection contains 24,000 training image samples, and 1000 test image samples. The image sample, as an aggregated entity, is defined as the entire set of RGB and depth images, which are acquired by all the installed data acquisition sensors at a particular instance of time. Furthermore, each image sample also includes data that relates to the detected 3D objects' spatial position and orientation, dimension and category.

The 3D objects that are defined and stored in the experimental dataset represent four main categories: pedestrians, cyclists, motorcyclists, and vehicles. These categories are represented in the data set with the weights 0.2, 0.2, 0.2, and 0.4, respectively. This ensures that the actual vehicles are assigned a proper weight, which corresponds to the real-world situation. The data curation phase ensures that the experimental dataset allows for each 3D object to manifest during eight image samples. This increases the structural diversity of the 3D objects that are stored in the experimental dataset, while allowing for a greater number of spatial positions and orientations to be stored. Additionally, it can be stated that the actual movement of the 3D objects considers the standard traffic rules.

The detection areas are determined by a rectangle with the dimensions of 100 by 50 m, in the case of the T-junction scenario. Additionally, in the context of the roundabout scenario, the detection areas are determined by squares with sides of 90 m. The area that is covered by the data acquisition sensors has a size of 4300 square metres in the case of the T-junction scenarios, and 12,200 square metres in the case of the roundabout scenarios. The algorithmic core of the proposed 3D objects detection model takes into consideration the mathematical model of a laser-based Lidar sensor, which is described in [11]. Additionally, further theoretical and practical aspects may be studied in paper [5,12].

5. The Training Process

The experimental dataset is processed through a training process considering the following steps. Thus, a particular instance of the 3D objects detection model is trained considering the data that is provided by multiple data acquisition sensors, and considering the algorithmic process that has already been described. The training process considers a Stochastic Gradient Descent (SGD) optimisation during 90 epochs. The learning rate is 10^{-4}, while the momentum is 0.8. Moreover, a loss function is considered, which penalises the regression relative to the position, size and yaw angle.

Considering the voxel size as (vox_a, vox_b, vox_c), then the anchor stride that goes along the dimensions X and Y, in the case of the T-junction, is set to (0.3, 0.3,0.5) m and 0.5 m, respectively. The consideration of exactly the same hyperparameters in the case of the roundabout is not feasible because the covered area is approximately three times larger. This would provoke computational problems as a consequence of the impossibility to store all the features maps in the graphics processing unit's (GPU) memory. Therefore, the spatial coverage of the axis X and Y is reduced, in the case of the roundabout, through a voxel size of (0.4, 0.4, 0.4) m, and an anchor stride of 0.8 m.

The algorithmic core, which actually performs the 3D object detection, is mostly designed to isolate vehicles from the processed image samples. The other three entity categories, pedestrian, cyclist, and motorcyclist, have the role of contributing to the prevention

of the statistical phenomenon of overfitting, since they allow for the trained model to learn the necessary distinct features for the vehicles.

Additionally, the proposed algorithmic core is designed to apply rotations to the determined bounding boxes. The angle of these rotations is selected, considering a randomized model, from the interval $[-26, 26]$. The rotation is applied in order to compute the rotation angle in an as general as possible way. Furthermore, the rotation also contributes to the prevention of the model reaching a state of overfitness.

6. Presentation of the Performance Assessment Process

The effective performance of the proposed 3D objects detection model is evaluated considering the two road scenarios, the T-junction and the roundabout. Additionally, the variation in the number of data acquisition sensors is assessed in connection to the accuracy of the 3D objects detection process itself.

6.1. Performance Assessment Metrics

The 3D objects detection process is assessed using four performance metrics. They are intersection relative to union (IRTU), recall, precision, and the communication cost, which is defined by the average volume of data that is transmitted between a data acquisition sensor and the central data processing components relative to each image sample. The communication cost is measured in kilobits. Let us recall that the concept of image sampling has already been described in a previous section.

The intersection relative to union essentially measures the spatial similarity of a pair of bounding boxes, one which is normally selected from the set of estimated bounding boxes, while the other is selected from the ground-truth set. This is determined by the following formula: $IRTU(B_{gt}, B_e) = \frac{volume(B_{gt} \cap B_e)}{volume(B_{gt} \cup B_e)}$. Here, B_{gt} and B_e represent the ground-truth and estimated bounding boxes, respectively. The set of estimated bounding boxes encompasses all the positive entities. These are the bounding boxes that are determined by the 3D objects detection algorithm considering a confidence score that is greater than a certain threshold, which is denoted by *Trs* in this paper. Let us recall that the metric IRTU also considers the location, size, and the yaw angle of both bounding boxes, which essentially represents the orientation. Thus, the value of this metric is 0, if the respective bounding boxes do not intersect, while it is 1 if the two bounding boxes are identical in terms of their size, orientation, and location. The value of the *Trs* has been calibrated through successive experimental trials. Thus, we have determined that a value of 0.75 generates an optimal balance between the quality of the 3D objects detection, and the utilized computational resources.

The precision is defined by the ratio between the number of estimated bounding boxes that are matched, considering the already mentioned definition, and the total number of bounding boxes that are part of the estimated set. Furthermore, the recall is determined by the ratio between the number of estimated bounding boxes that are matched relative to the overall number of bounding boxes that are part of the ground-truth set. It is natural to observe that precision and recall are, in essence, functions of *Trs*. The arithmetic and computational relationship between precision and recall is studied in the related scientific literature. The interested reader may find further details in paper [26].

6.2. Comparative Evaluation Metrics of Several Fusion Schemes

Considering the fusion schemes that are presented in the introductory part of this paper, the purpose of this experiment is to compare the performance of early fusion (EF) and late fusion (LF) schemes considering the effective detection performance, and also the communication cost and computational time. The evaluation considers both road topologies, the T-junction and the roundabout. Considering the late fusion scheme, the algorithm uses an IRTU threshold *Trs* with a value of 0.17. This value has been experimentally calibrated in order to prevent the same 3D object being detected multiple times. The values of the performance metrics can be studied in Table 2. Let us note that the

communication cost is computed for each sensor, while the computation time is calculated for each image sample. Let us recall that the concept of image sample has already been defined in a previous section.

Table 2. Comparative performance analysis between late and early fusion schemes.

Scheme and Topology	Communication Cost (Kilobits)	Computation Time (ms)
LF T-junction	0.39	217
EF T-junction	471	296
LF Roundabout	0.17	139
EF Roundabout	541	228

The experimental results that are presented in Table 2 suggest that the more efficient 3D object detection in the case of the early fusion schemes comes at the expense of a greater computational cost. This behaviour is determined by the larger data volume that is transmitted in order to send the raw point clouds from the data acquisition sensors to the central processing components, as compared to the similar data transmission in the case of the late fusion. Furthermore, the experimental results that were obtained demonstrate that the early fusion variant exhibits a higher performance concerning the 3D objects detection in the case when the threshold Trs is assigned higher values. The best performance was obtained for $Trs = 0.92$. Furthermore, considering a particular value of Trs, the 3D object detection performance is more efficient in the case of the T-junction road topology, as compared to the roundabout topology. This behaviour is determined by the larger voxels that have to be computed in the roundabout scenario, as it has already been explained in a previous section.

Additionally, the results that are presented in Table 2 suggest that the early fusion variant determines a higher communication cost. The explanation resides in the larger volume of data that has to be transmitted in order to accommodate the raw point clouds. Furthermore, it should be observed that the required capacity of the data transmission link is dependent on the frequency of the processed image samples. As an example, considering a frequency of ten image samples per second, which is common in the case of laser-based Lidar sensors, the required capacity of the data transmission link is 4.71 Mbps. This number is obtained through the multiplication of the communication cost value by 10, considering there are 10 image samples that are processed in a second. The field experiments that were conducted demonstrate that this transmission rate can be easily accommodated by the wireless or wired data connections that are available along the road that is monitored by the data acquisition sensors. Moreover, considering that the temporal sequence of the transmitted image samples is relevant, then it can be stated that the network latency may produce problems concerning the actual 3D object detection. The field studies that were conducted did not detect this problem, and we are able to make the assumption that this problem may occur only in cases when the latency of the data link is greater that a few tens of milliseconds. Nevertheless, a systematic study of the latency in the context of automotive data transmission may constitute the object of a separate paper.

The computational time that is necessary in order to process each image sample is greater in the context of the early fusion schemes, as there are more point clouds that have to be processed than in the case of the late fusion scheme. Furthermore, the computational time depends on the GPU that is used in order to deploy the central data processing components. The experiments that we conducted considered an Nvidia RTX 3090 GPU. The hardware components that sustain the function of the central data processing assembly are installed in a mobile van for the purpose of this experimental process. The data acquisition sensors are capable of sending the collected data to the central data processing components using a wired or wireless data connection. The experimental road deployments that are reported in this paper consider wireless connections through the 802.11 family of standards. The system may be easily modified in order to support wireless standards that offer a longer range. Additionally, it is relevant to note that in the case of permanent road deployments,

it is possible to configure and deploy wired data connections between the data acquisition sensors and the central data processing components. The wired data connections may use either standard electrical conductors, or fiber-optic cables. It is important to note that the hardware features of the data transmission infrastructure can be easily and transparently modified relative to the deployed 3D objects detection system.

6.3. Investigation Concerning the Number and Placement of Sensors

This stage of the experimental process considers the assessment of the impact that the number, spatial position and orientation of the data acquisition sensors have on the actual detection of the 3D objects. This evaluation considers the early and late fusion schemes. The same structure of the algorithmic core is considered for the actual 3D objects detection. In Table 3, the detection performance is measured through the actual accuracy of the detected 3D objects. The accuracy quantifies the number of precise 3D object detections relative to the total number of 3D objects that are part of the experiment.

Table 3. Comparative performance analysis considering various numbers of the active sensors.

No. Sensors	T-Junction EF	T-Junction LF	Roundabout EF	Roundabout LF
1	0.212	0.186	0.196	0.174
2	0.243	0.205	0.207	0.198
3	0.324	0.308	0.305	0.289
4	0.439	0.413	0.424	0.402
5	0.547	0.532	0.536	0.514
6	0.657	0.635	0.638	0.618
7	0.789	0.768	0.778	0.769
8	0.858	0.837	0.849	0.828
9	0.957	0.946	0.948	0.939
10	0.992	0.989	0.978	0.968
11	–	–	0.989	0.985
12	–	–	0.994	0.992

The values of the performance metric, which are presented in Table 3, demonstrate that the accuracy of the 3D objects detection improves directly proportional to the numbers of active data acquisition sensors. It is interesting to note that, in the case of the T-junction topology, the optimal detection accuracy is reached with ten active data acquisition sensors, while the optimal detection accuracy relative to the roundabout topology is achieved when twelve active data acquisition sensors are considered. The experiments that were performed prove that the optimal level of the accuracy is greater or equal than 0.98. This implies that the 3D objects are detected without any significant issues, and the autonomous driving process occurs in an adequate manner. Additionally, more data acquisition sensors are required, if the detection accuracy level is necessary to be maintained on a larger area that is monitored. Let us recall that the monitored surface area is 4300 square metres in the case of the T-junction scenario, while the roundabout scenario covers an area with a surface of 12,200 square metres. Moreover, it can be observed that the early fusion scheme determines a superior level of the accuracy considering all the evaluated cases. This is a direct consequence of the early fusion scheme's ability to use more data during the preprocessing stage, as compared to the late fusion detection model.

Furthermore, the experimental workflow considers the spatial position and orientation of the data acquisition sensors. This is particularly important in the case of the T-junction scenario, which benefits from groups of three data acqusition sensors that are deployed in order to monitor certain sub-sections of the overall detection area. The next paragraph discusses on the experimental results that were obtained during the assessment of the spatial diversity's impact on the actual 3D objects detection accuracy.

The experiments that were conducted demonstrate that the problem of the data acquisition sensors' spatial diversity [27] is also important. Thus, the prevention or, at least, minimization of the multihop data acquisition links from the sensors to the central data

processing components is required in order to obtain a high level of the 3D object detection accuracy. Thus, we have determined that, in the case of the T-junction scenario, a group of two sensors outperforms the single most efficient data acquisition sensor by 57%, while the optimal cluster of three sensors, which monitor precise sub-sections of the overall area, determine an improvement of 96%, relative to the most efficient data acquisition sensor. Furthermore, the same performance gains are 46% and 82%, respectively, in the case of the roundabout scenario. The percentage of the improvement is calculated relative to the base value of the accuracy, as it is determined by an individual data acquisition sensor considering the roundabout and T-junction scenarios. This demonstrates that the clusters of sensors should provide optimal overlap between their members. Consequently, this further demonstrates that the early fusion scheme may reduce the occurrence of falsely detected 3D objects. Additionally, clusters of sensors with properly overlapped members contribute to the increase of the 3D object detection accuracy.

6.4. Comparative Performance Analysis Relative to MV3D, AVOD and F-PointNet

Let us recall that the described 3D object detection model represents a significant extension and optimization of the F-PointNet fusion based method. Therefore, we have performed a comparative performance analysis with the reference F-PointNet model, and also relative to the MV3D and AVOD models. We used the same dataset that has been generated by the experimental field setup that has already been described. Furthermore, we implemented the MV3D, AVOD and F-PointNet fusion based models, as it is suggested in papers [2–4], respectively. Following this, the detection performance is computed considering the same road scenarios and number of sensors, which are presented in Table 3. Let us recall that the detection performance is measured through the actual accuracy of the detected 3D objects. The accuracy quantifies the number of precise 3D object detections relative to the total number of 3D objects that are part of the experiment. Thus, let us study in Table 4 the detection performance that is determined by the MV3D model.

Table 4. Comparative performance analysis considering the MV3D model and various numbers of the active sensors.

No. Sensors	T-Junction EF	T-Junction LF	Roundabout EF	Roundabout LF
1	0.122	0.114	0.119	0.108
2	0.173	0.178	0.185	0.142
3	0.215	0.256	0.277	0.203
4	0.289	0.356	0.352	0.346
5	0.398	0.431	0.468	0.424
6	0.508	0.513	0.528	0.505
7	0.623	0.597	0.599	0.594
8	0.698	0.684	0.675	0.649
9	0.789	0.768	0.759	0.702
10	0.841	0.869	0.854	0.789
11	–	–	0.869	0.828
12	–	–	0.904	0.893

Following, let us study in Table 5 the detection performance that is determined by the AVOD model.

Table 5. Comparative performance analysis considering the AVOD model and various numbers of the active sensors.

No. Sensors	T-Junction EF	T-Junction LF	Roundabout EF	Roundabout LF
1	0.137	0.132	0.133	0.119
2	0.192	0.188	0.194	0.158
3	0.228	0.269	0.290	0.213
4	0.304	0.372	0.371	0.358
5	0.407	0.441	0.479	0.443
6	0.517	0.531	0.540	0.520
7	0.639	0.604	0.608	0.608
8	0.708	0.697	0.688	0.668
9	0.804	0.779	0.767	0.717
10	0.861	0.879	0.867	0.799
11	–	–	0.878	0.845
12	–	–	0.924	0.902

Finally, let us observe in Table 6 the detection performance that is determined by the F-PointNet model.

Table 6. Comparative performance analysis considering the F-PointNet model and various numbers of the active sensors.

No. Sensors	T-Junction EF	T-Junction LF	Roundabout EF	Roundabout LF
1	0.151	0.141	0.148	0.134
2	0.199	0.206	0.217	0.163
3	0.241	0.293	0.305	0.231
4	0.316	0.386	0.391	0.368
5	0.416	0.469	0.490	0.452
6	0.540	0.542	0.551	0.539
7	0.641	0.628	0.610	0.608
8	0.716	0.709	0.692	0.668
9	0.831	0.789	0.776	0.730
10	0.878	0.898	0.881	0.804
11	–	–	0.899	0.846
12	–	–	0.926	0.907

The comparative performance data shows that there are marginal differences regarding the detection accuracy between the three fusion based reference models. Furthermore, this comparative experimental evaluation demonstrates that the integrated 3D object detection system that is reported in this paper, which is featured by an improved 3D object detection core, and an original architectural structure of its software components, determines a superior 3D object detection accuracy compared to all the reference fusion based schemes.

6.5. Remarks Regarding Additional Road Traffic Scenarios

The described system has been deployed to another national road, apart from that already described, considering both the roundabout and T-junction scenario. The experiment assessed the deployed system's behaviour over a period of one month. The system has been tested in various weather conditions, including thick fog and heavy rain. Furthermore, the density of the installed sensors has been progressively reduced. The results of the experimental process are presented through the detection performance values, which are displayed in Table 7. Let us recall that the detection performance is determined by the number of precise 3D object detections relative to the total number of 3D objects that are part of the experiment.

Table 7. Outcome of the supplementary assessment process.

No. Sensors	T-Junction Fog	T-Junction Rain	Roundabout Fog	Roundabout Rain
4	0.436	0.412	0.422	0.401
6	0.655	0.634	0.635	0.616
8	0.854	0.835	0.846	0.825
10	0.988	0.983	0.972	0.964

The data that is described in Table 7 demonstrates that the system is not substantially influenced by the weather conditions. Furthermore, it can be observed that the system provides acceptable detection performance in the case when less detection sensors are deployed. Thus, the outcomes of these initial experimental assessment processes suggest that the system can be considered as a sufficiently economical solution for the implementation of an effective autonomous driving approach. In Figure 2, the placement of the sensors is marked with a red arrow considering the T-junction area. Thus, it can be easily observed that six sensors are enough in order to sufficiently cover the area of the T-junction proper. The additional data acquisition sensors, when they are available, can be used in order to perform the pre-calibration of the system prior to the autonomous vehicle arrival in the zone of the T-junction.

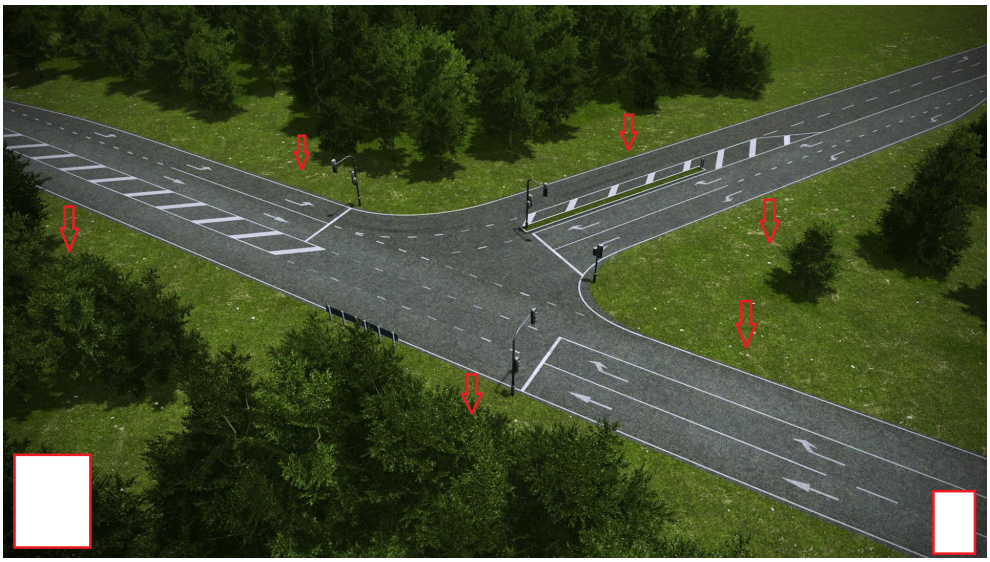

Figure 2. Sensors placement in the immediate vicinity of the T-junction.

Additionally, in Figure 3, the placement of the sensors is also marked with a red arrow considering the roundabout area. Similarly to the T-junction topology, it can be easily observed that eight sensors are enough in order to efficiently cover the area of the T-junction proper. The experimental assessment that was performed suggests that sufficient levels for the detection accuracy can be obtained with only six sensors. The additional data acquisition sensors, when they are available, may be used in order to perform the pre-calibration of the system prior to the autonomous vehicle arrival in the area of the roundabout.

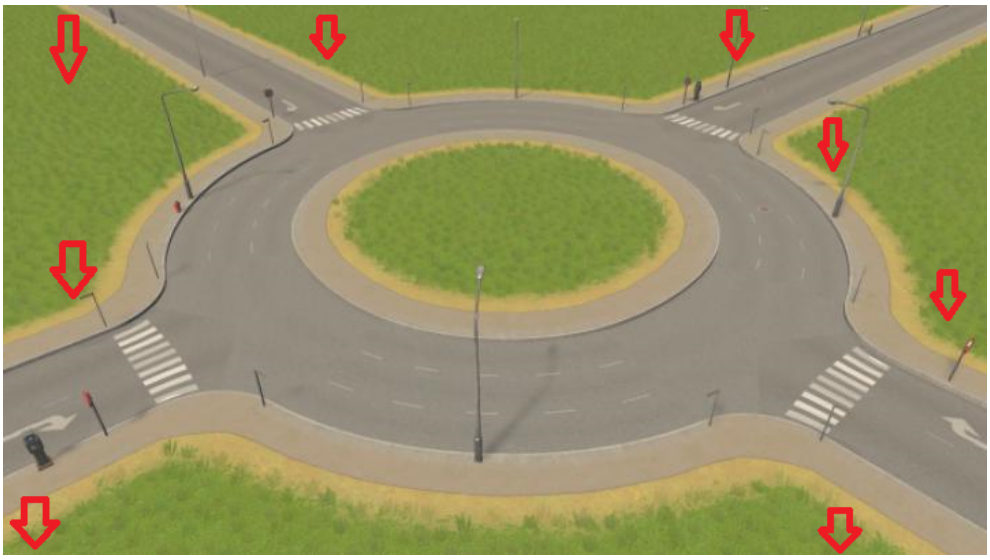

Figure 3. Sensors placement in the immediate vicinity of the roundabout.

7. Conclusions and Future Work

This paper describes a comprehensive study that proposes an improved integrated approach, which may enhance the functional accuracy, reliability and safety of autonomously driven vehicles. Essentially, the system that is described considers early and late fusion data processing and detection schemes. The system is based on the utilization of data acquisition sensors, which are mounted on masts at the side of the road. The data that are gathered by the sensors are aggregated and sent over to the central data processing components. The generated bounding boxes are shared with the vehicles that drive in the monitored zone. The proposed system's validity and effectiveness are tested considering two relatively difficult road topologies, a T-junction and a roundabout. The results of the experimental evaluation suggest that an increased number of data acquisition sensors is necessary in order to obtain image samples that can be effectively used during the 3D object detection process. In essence, the experiments that were conducted demonstrate that the described system is able to significantly reduce the rate of incorrectly detected 3D objects, while improving the overall accuracy of the detection process. Considering the experimental results that were presented, it can be stated that the system is able to accurately process the image samples and detect the 3D objects in virtually all of the cases. Additionally, it is important to note that the actual hardware deployment of the system considers existing technologies and data transmission protocols. This minimizes the implementation costs, while the data sharing from the central data processing components to the individual cars in the monitored area ensures that even the vehicles that are not equipped with the latest technologies may benefit from the most efficient autonomous driving experience. The integrated system's 3D object detection accuracy is assessed against the reference fusion based models MV3D, AVOD and F-PointNet. The results prove that the enhanced detection core and architecture of the described integrated system generates a higher level of performance relative to all three reference fusion based models. The field research setup that generated the experimental data that we considered, which has been deployed with the support of our industry partners, is already supporting the efforts of the relevant car manufacturers' research teams that aim to improve the current autonomous driving approaches.

Therefore, the described system will be enhanced in several respects. Thus, the implementation of the early and late fusion schemes will be improved, with an emphasis on the early fusion schemes, which exhibit the most efficient computational behaviour in practice. Additionally, the field placement of the data acquisition sensors should be improved through successive empirical trials. The goal is to obtain the greatest possible coverage with the smallest number of data acquisition sensors, while maintaining the high levels of detection accuracy that are reported in this paper. Furthermore, the relevant re-

search requirements that are issued by our automotive industry partners will be considered during the upcoming development process of the integrated 3D object detection system.

Author Contributions: Conceptualization, R.B. and D.B.; methodology, R.B., D.B. and M.I.; software, R.B.; validation, R.B., D.B. and M.I.; formal analysis, D.B. and M.I.; investigation, R.B.; resources, R.B., D.B. and M.I.; data curation, R.B.; writing—R.B.; writing—review and editing, R.B., D.B. and M.I.; supervision, R.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Geiger, A.; Lenz, P.; Urtasun, R. Are We Ready for Autonomous Driving? The KITTI vision benchmark suite. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012; Volume 2012; pp. 3354–3361.
2. Chen, X.; Ma, H.; Wan, J.; Li, B.; Xia, T. Multi-View 3D Object Detection Network for Autonomous Driving. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
3. Ku, J.; Mozifian, M.; Lee, J.; Harakeh, A.; Waslander, S. Joint 3D proposal generation and object detection from view aggregation. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018.
4. Qi, C.R.; Liu, W.; Wu, C.; Su, H.; Guibas, L.J. Frustum PointNets for 3D Object Detection From RGB-D Data. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018 .
5. Ortiz, L.E.; Cabrera, E.V.; Gonçalves, L.M. Depth data error modeling of the zed 3D vision sensor from stereolabs. *ELCVIA Electron. Lett. Comput. Vis. Image Anal.* **2013**, *17*, 1–15. [CrossRef]
6. Roberts, R.; Sinha, S.N.; Szeliski, R.; Steedly, D. Structure from motion for scenes with large duplicate structures. In Proceedings of the CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011; pp. 3137–3144.
7. Arnold, E.; Al-Jarrah, O.Y.; Dianati, M.; Fallah, S.; Oxtoby, D.; Mouzakitis, A. A Survey on 3D Object Detection Methods for Autonomous Driving Applications. *IEEE Trans. Intell. Transp. Syst.* **2019**, *20*, 3782–3795. [CrossRef]
8. Beltran, J.; Guindel, C.; Moreno, F.M.; Cruzado, D.; Garcia, F.; De La Escalera, A. Birdnet: A 3D object detection framework from Lidar information. In Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, USA, 4–7 November 2018; pp. 3517–3523.
9. Li, B.; Zhang, T.; Xia, T. Vehicle Detection from 3D Lidar Using Fully Convolutional Network. In Proceedings of the Robotics: Science and Systems, Ann Arbor, MI, USA, 18–22 June 2016.
10. Simony, M.; Milzy, S.; Amendey, K.; Gross, H.-M. Complex-yolo: An euler-region-proposal for real-time 3D object detection on point clouds. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018.
11. Castanedo, F. A review of data fusion techniques. *Sci. World J.* **2013**, *2013*, 142–149. [CrossRef] [PubMed]
12. Feng, D.; Haase-Schuetz, C.; Rosenbaum, L.; Hertlein, H.; Duffhaus, F.; Glaeser, C.; Wiesbeck, W.; Dietmayer, K. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *arXiv* **2019**, arXiv:1902.07830.
13. Zhou, Y.; Tuzel, O. Voxelnet: End-to-end learning for point cloud based 3D object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
14. Shi, S.; Wang, X.; Li, H. Pointrcnn: 3D object proposal generation and detection from point cloud. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
15. Yang, S.; Sun, Y.; Liu, S.; Shen, X.; Jia, J. Std: Sparse-to-dense 3D object detector for point cloud. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 1951–1960.
16. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. PointNet: Deep learning on point sets for 3D classification and segmentation. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 652–660.
17. Chen, Q.; Tang, S.; Yang, Q.; Fu, S. Cooper: Cooperative Perception for Connected Autonomous Vehicles based on 3D Point Clouds. In Proceedings of the 39th IEEE International Conference on Distributed Computing Systems (ICDCS), Dallas, TX, USA, 7–10 July 2019.
18. Chen, Q.; Ma, X.; Tang, S.; Guo, J.; Yang, Q.; Fu, S. F-Cooper: Feature based Cooperative Perception for Autonomous Vehicle Edge Computing System Using 3D Point Clouds. In Proceedings of the IEEE/ACM Symposium on Edge Computing (SEC), Washington, DC, USA, 7–9 November 2019.

19. Hurl, B.; Kohen, R.; Czarnecki, K.; Waslander, S. Trupercept: Trust modelling for autonomous vehicle cooperative perception from synthetic data. In Proceedings of the 2020 IEEE Intelligent Vehicles Symposium (IV), Las Vegas, NV, USA, 19 October–13 November 2020.
20. Ghamisi, P.; Rasti, B.; Yokoya, N.; Wang, Q.; Hofle, B.; Bruzzone, L.; Bovolo, F.; Chi, M.; Anders, K.; Gloaguen, R. Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art. *IEEE Geosci. Remote Sens. Mag.* **2019**, *7*, 6–39. [CrossRef]
21. Yin, L.; Wang, X.; Ni, Y.; Zhou, K.; Zhang, J. Extrinsic parameters calibration method of cameras with non-overlapping fields of view in airborne remote sensing. *Remote Sens.* **2018**, *10*, 1298. [CrossRef]
22. Yue, R.; Xu, H.; Wu, J.; Sun, R.; Yuan, C. Data registration with ground points for roadside Lidar sensors. *Remote Sens.* **2019**, *11*, 1354. [CrossRef]
23. Knorr, M.; Niehsen, W.; Stiller, C. Online extrinsic multi-camera calibration using ground plane induced homographies. In Proceedings of the 2013 IEEE Intelligent Vehicles Symposium (IV), Gold Coast, Australia, 23–26 June 2013; pp. 236–241.
24. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Realtime Object Detection with Region Proposal Networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 91–99.
25. Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; Coltun, V. CARLA: An open urban driving simulator. In Proceedings of the 1st Annual Conference on Robot Learning, Mountain View, CA, USA, 13–15 November 2017; pp. 1–16.
26. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. *Int. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
27. Xu, S.; Liu, H.; Gao, F.; Wang, Z. Compressive Sensing Based Radio Tomographic Imaging with Spatial Diversity. *Sensors* **2019**, *19*, 439. [CrossRef] [PubMed]
28. Schlosser, J.; Chow, C.K.; Kira, Z. Fusing Lidar and images for pedestrian detection using convolutional neural networks. In Proceedings of the 2016 IEEE International Conference on Robotics and Automation (ICRA), Stockholm, Sweden, 16–21 May 2016; pp. 2198–2205.
29. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2015**, arXiv:1409.1556.
30. Du, X.; Ang, M.H.; Karaman, S.; Rus, D. A general pipeline for 3D detection of vehicles. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation, Brisbane, Australia, 21–25 May 2018; pp. 3194–3200.
31. Lin, T.-Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.