future internet

MDPI

Article

NeXtFusion: Attention-Based Camera-Radar Fusion Network for Improved Three-Dimensional Object Detection and Tracking

Priyank Kalgaonkar and Mohamed El-Sharkawy *

Department of Electrical and Computer Engineering, Purdue School of Engineering and Technology, Indianapolis, IN 46202, USA; pkalgaon@purdue.edu
* Correspondence: melshark@purdue.edu

Abstract: Accurate perception is crucial for autonomous vehicles (AVs) to navigate safely, especially in adverse weather and lighting conditions where single-sensor networks (e.g., cameras or radar) struggle with reduced maneuverability and unrecognizable targets. Deep Camera-Radar fusion neural networks offer a promising solution for reliable AV perception under any weather and lighting conditions. Cameras provide rich semantic information, while radars act like an X-ray vision, piercing through fog and darkness. This work proposes a novel, efficient Camera-Radar fusion network called NeXtFusion for robust AV perception with an improvement in object detection accuracy and tracking. Our proposed approach of utilizing an attention module enhances crucial feature representation for object detection while minimizing information loss from multi-modal data. Extensive experiments on the challenging nuScenes dataset demonstrate NeXtFusion's superior performance in detecting small and distant objects compared to other methods. Notably, NeXtFusion achieves the highest mAP score (0.473) on the nuScenes validation set, outperforming competitors like OFT (35.1% improvement) and MonoDIS (9.5% improvement). Additionally, NeXtFusion demonstrates strong performance in other metrics like mATE (0.449) and mAOE (0.534), highlighting its overall effectiveness in 3D object detection. Furthermore, visualizations of nuScenes data processed by NeXtFusion further demonstrate its capability to handle diverse real-world scenarios. These results suggest that NeXtFusion is a promising deep fusion network for improving AV perception and safety for autonomous driving.

Keywords: CondenseNeXt; sensor fusion; object detection; autonomous vehicle; PyTorch

Citation: Kalgaonkar, P.; El-Sharkawy, M. NeXtFusion: Attention-Based Camera-Radar Fusion Network for Improved Three-Dimensional Object Detection and Tracking. *Future Internet* **2024**, *16*, 114. https://doi.org/10.3390/fi16040114

Academic Editors: Eirini Eleni Tsiropoulou and Symeon Papavassiliou

Received: 19 February 2024
Revised: 23 March 2024
Accepted: 27 March 2024
Published: 28 March 2024

1. Introduction

There has been a rapid advancement in the development of sensing systems for autonomous driving that has notably elevated the effectiveness of perception tasks, such as object detection, in recent years. Despite these achievements in research and development, there remains a lack of widespread adoption of level 4 or 5 autonomous driving capabilities in commercial vehicles due to autonomous vehicles (AVs)' reliance on single-sensor perception of the real world and a substantial commitment to research and development to guarantee the continuous evolution and enhancement in the technology over time [1]. Furthermore, the prediction and decision—making processes in AVs that rely on a single sensor can be hampered by external factors such as bad weather, occlusion, or poor lighting conditions because cameras struggle in low-light environments, whereas radars cannot detect objects with rich visual features. This limitation in cameras and radar, and the potential consequences of reliance on a single sensor for object detection in AVs, has generated significant attention in the field of research toward the utilization of multi-modal based sensing in the automotive domain, especially in perception systems that fuse camera and radar inputs [2–4].

An ideal fusion system utilizing both the cameras and radar sensor information can effectively leverage the advantages of both these sensors, concurrently addressing the limitations inherent in each. While a camera as a sensor input offers detailed texture and

semantic information, its performance diminishes with long-range small objects, occlusions, and poor lighting conditions, whereas radar as a sensor input exhibits the ability to offer reliable performance in all weather and lighting conditions, detect small objects at long ranges, and operate without hindrance from issues related to occlusions. However, radar sensors encounter difficulties in precisely identifying objects due to the absence of detailed texture and semantic features [2,5,6]. The work presented within this paper revolves around the primary objective of determining how we can effectively harness the benefits of both modalities (camera and radar sensors) to attain precise and dependable object detection.

An optimal Camera-Radar fusion network must capitalize on the advantages offered by both sensors. Simultaneously, it should also guarantee that the limitations of one sensor do not impact the performance of the other. Previous studies in the fusion of camera and radar modalities have often employed the mapping of radar data onto the camera's data [7]. However, working within this technique imposes limitations on performance, particularly in scenarios involving object occlusion, thus resulting in inefficient use of radar sensor data. More sophisticated and cutting-edge methodologies engage in fusion at the feature level instead of directly mapping features. For instance, the approach proposed for the AVOD network [8] extracts bird-eye view features concurrently from the camera and the radar's sensor input. This approach then merges these features on a per-object basis to capitalize on the unique information extracted by each of these modalities to perform camera-radar sensor fusion. However, it is observed that the concurrent approach of extracting features and performing fusion does not hold true for instances where the camera sensor becomes unreliable in situations such as occluded objects or adverse conditions, like rain or fog. In such scenarios, radar-based sensors are not affected and work well, but the reliability of camera-based sensors can significantly decrease, resulting in a notable overall performance decline in the AV system for object detection in such adverse conditions.

Evidently, there is a necessity to enhance the dependability of camera-radar systems to attain satisfactory performance, particularly in situations where the quality of camera input is compromised due to external factors. The multi-modal sensor fusion network proposed within this paper posits that by independently extracting valuable information from both camera and radar sensors, one can leverage the advantages of each modality without compromising either in situations of degraded performance due to external factors. This innovative multi-modal fusion approach is rooted in the acknowledgment that cameras and radars offer complementary attributes. The detailed texture and semantic-rich information from cameras can be utilized to identify multiple different objects, while radars offer the advantage of detecting objects over long distances, unobstructed by occlusion, and are reliable in adverse weather conditions such as fog or rain. Thus, this potential to individually extract information serves as a method to enhance the overall reliability of the proposed neural network for 3D object detection and tracking names as NeXtFusion, proposed within this paper.

NeXtFusion network utilizes point-cloud data representation for object detection, extracting all information obtained from radar sensor(s), whereas the semantic information obtained from camera sensor(s) primarily serves to distinctly identify objects within an input image/video. Drawing inspiration from this paper's authors' most recent work, NeXtDet [9], significant advancements are being proposed in this paper where the literature on camera-based semantic information extraction is relied upon to independently and efficiently extract these semantic-rich features from camera-based RGB images and fuse them with radar's point-cloud data as illustrated in Figure 1. The structure of this paper is organized as follows: Section 2 offers the reader information on related literature and studies. Section 3 provides information about the methodology for the proposed multi-modal sensor-fusion 3D object detection model called NeXtFusion. Section 4 assesses the performance of the proposed network through benchmarking on a multi-modal large-scale autonomous driving dataset and visualizing the results. Finally, Section 5 presents the concluding remarks for this paper.

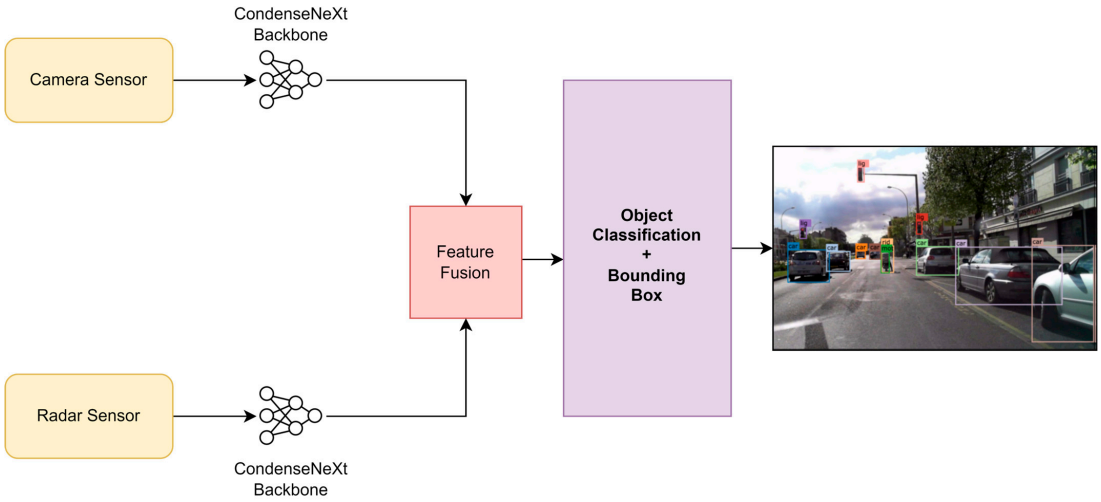

Figure 1. An overview of the proposed NeXtFusion multi-modal sensor fusion neural network.

2. Literature Review on Relevant Studies

2.1. Camera-Based Object Detection

A convolutional neural network (CNN) represents a subset of the broader deep neural network (DNN) extensively utilized in the realm of AI. It is frequently employed for devising inventive methodologies and algorithms within the OpenCV realm, facilitating complex tasks such as image classification and object detection. This is achieved through the utilization of multiple layers of neurons, simulating the natural visual perception of human beings [10]. Object detection involves a computer vision process aimed at identifying instances of objects belonging to a specific class, categorizing the types of objects, pinpointing their locations, and precisely labeling them within an input image or video.

A state-of-the-art object detector essentially pinpoints the position and type of objects within an image [9]. It is a system with three main modules (parts):

1. Backbone: This module acts as the foundation, extracting salient features from the image and producing a compressed representation through a robust image classifier. Imagine it like a skilled photographer capturing the essence of a scene.
2. Neck: This module acts as the bridge, connecting the backbone and skillfully merging features extracted from different levels of the backbone. Consider it like a data sculptor, gathering and harmonizing different perspectives to create a richer understanding.
3. Head: This module is the decision maker responsible for drawing bounding boxes around objects and classifying their types. Think of it as a detective, analyzing the combined information and identifying what each object is and where it lies.

Modern object detectors employ a specialized component known as the head, tasked with both object localization and classification, as aforementioned. Two predominant approaches prevail in head design: single-stage and two-stage architectures.

Single-stage detectors, epitomized by YOLO [11], SSD [12], RetinaNet [13], CornerNet [14], and CenterNet [15], prioritize speed and efficiency. They execute both tasks—generating bounding boxes and classifying object types—within a single, unified module. This streamlined approach enables faster inference, making them attractive for real-time applications on computationally constrained platforms like edge devices. However, their focus on efficiency might compromise detection accuracy compared to their two-stage counterparts. In contrast, two-stage detectors, exemplified by the R-CNN family (including Fast R-CNN [16], Faster R-CNN [17], Mask-RCNN [18], Cascade R-CNN [19], and Libra R-CNN [20]), prioritize accuracy over speed. These detectors follow a two-step process: First, a dedicated region proposal network (RPN) generates potential object regions within the image. Subsequently, these proposed regions are forwarded to a separate module for classification and precise bounding box refinement. This

division of labor leads to superior detection performance but incurs a higher computational cost, limiting their real-time capabilities.

When choosing an object detector, striking a balance between accuracy and inference speed is crucial. single-stage detectors like YOLO offer impressive speed, making them suitable for real-time applications like autonomous vehicles. However, for tasks demanding high accuracy, two-stage detectors like Faster R-CNN might be preferred despite their slower performance [21]. Ultimately, the optimal choice hinges on the specific requirements and resource constraints of the application.

The neck of an object detector acts as a bridge, meticulously combining features extracted from various depths within the backbone. This intricate network of interconnected pathways, both descending (top-down) and ascending (bottom-up), allows for the seamless integration of multi-scale information. Popular strategies employed within the neck include the addition of specialized modules, like spatial pyramid pooling [22], which further enhance feature fusion capabilities. Alternatively, path-aggregation blocks such as feature pyramid networks [23] or path-aggregation networks [24] may be employed. The backbone architecture typically relies on a resilient image classifier for implementation, such as CondenseNeXt [25]. It serves as an effective and resilient image classification network developed to efficiently utilize the computational resources needed for real-time inference on edge devices with limited processing power. The CondenseNeXt CNN is employed in the NeXtDet object detection network as the backbone and has been further modified to achieve even greater lightweight characteristics than its original design for image classification purposes by deprecating the final layers of classification to ensure compatibility and facilitating linking the backbone to the neck module of NeXtDet. Thus, the NeXtDet network will serve as the foundational reference point for the multi-modal sensor-fusion research and enhancements presented within this paper.

2.2. Radar Point Cloud and Camera-Based Sensor Fusion for Object Detection

Radar sensors actively perceive the surroundings and analyze the reflected waves to determine the position and the speed of objects by constantly emitting radio waves. The world of radar signal processing for autonomous vehicles offers a rich tapestry of techniques to extract meaningful information from the sensor's output, for example, doppler processing [26], occupancy grid maps [27], multi-input multi-output (MIMO) radar [28], and representations using radar point clouds [29]. The work presented in this paper focuses on radar point-cloud data representation for radar signal processing using the proposed deep neural network for 3D object detection.

Typically, radars in automotive applications determine objects as 2D reference points in bird's eye view (BEV), synthesizing information on the radial distance to/from the object. In each radar detection, information is provided about the object's instantaneous velocity in the radial direction, along with the azimuth angle. The azimuth angle represents the horizontal angle of the object concerning the observer, which, in this case, is the radar sensor. The azimuth angle is crucial for determining the location of detected objects in the radar's field of view. Figure 2 provides a graphical representation of the difference between the radial velocity and the true velocity.

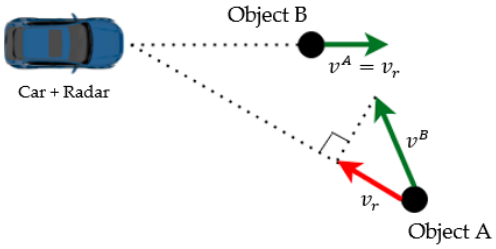

Figure 2. Comparison between velocities: the true velocity (v^A) for object A is equal to the radial velocity (v_r), whereas, for object B, the radar-reported radial velocity (v_r) indicated by the red arrow is not equal to the true velocity (v^B).

3. The Proposed NeXtFusion 3D Object Detection Network

This paper introduces NeXtFusion, a novel approach to 3D object detection for autonomous vehicles. It leverages sensor fusion, combining camera and radar data within a unified framework. By building upon the efficient NeXtDet framework [9], NeXtFusion prioritizes both performance and resource efficiency, making it particularly suitable for real-time applications. The key contribution of the work presented in this paper is in adapting the NeXtDet network for multi-modal sensor fusion and significantly improving 3D object detection through a lightweight object detection architecture. Section 4 provides details about extensive nuScenes experiments performed in order to demonstrate the proposed NeXtFusion network's ability to surpass existing benchmarks. These results pave the way for more robust and resource-conscious object detection for fully autonomous driving applications. Figure 3 illustrates the NeXtDet architecture that serves as the foundation for the proposed NeXtFusion network.

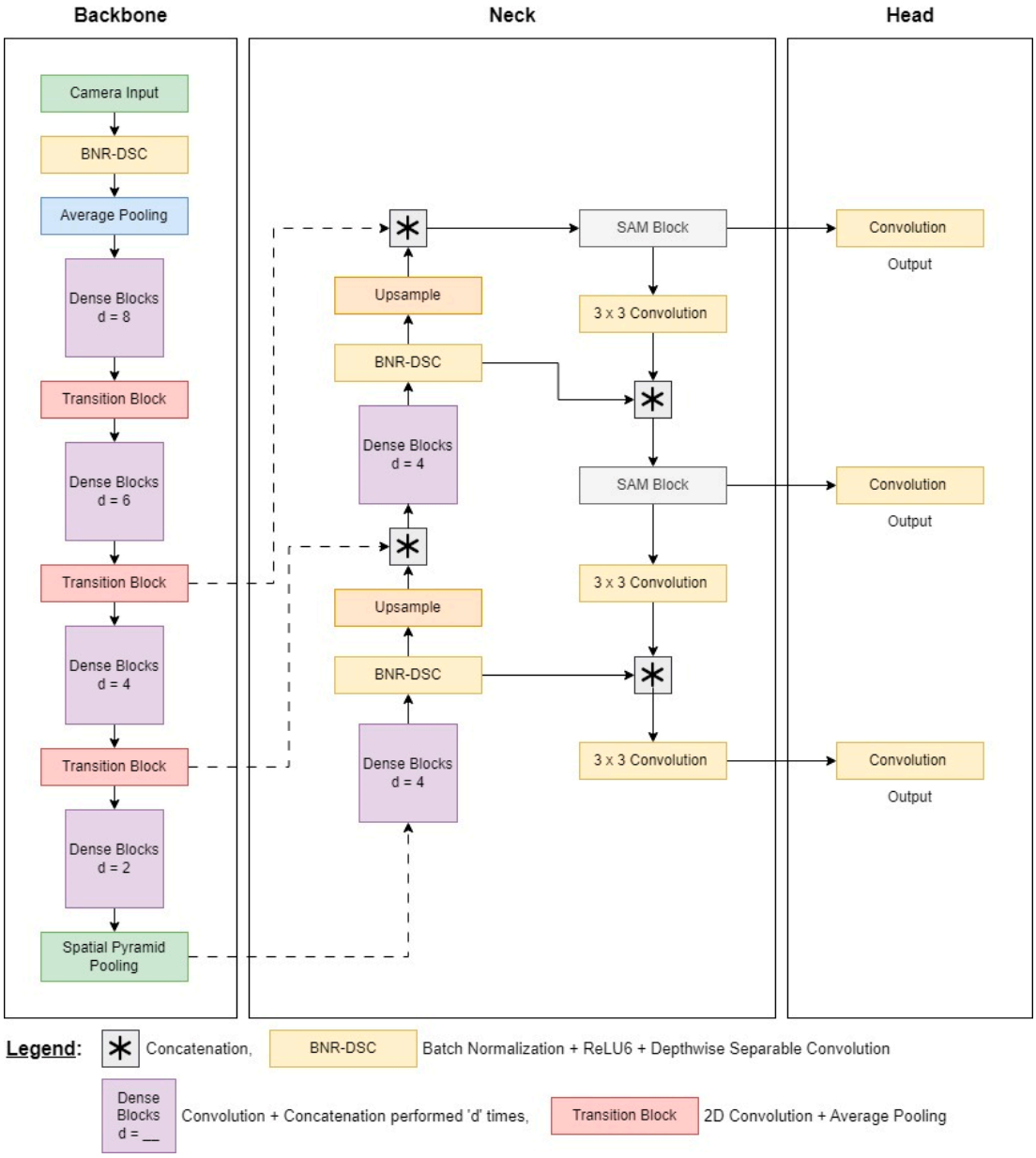

Figure 3. NextDet architecture utilized as the baseline for the proposed NeXtFusion architecture.

3.1. Backbone

The core of any cutting-edge object classifier is a powerful image-processing module called the backbone. This module scans an input image and distills features at different depths (levels of detail). Backbones typically rely on a resilient image classifier for implementation, which serves as an effective and resilient image classification network developed to efficiently utilize the computational resources needed for real-time inference on edge devices with limited processing power. NeXtDet utilizes CondenseNeXt CNN [25] in its backbone module, which also serves as the backbone for the proposed NeXtFusion network.

CondenseNeXt belongs to the DenseNet [30] family. It utilizes an innovative approach to capture spatial details from individual layers and transmit them in a feed-forward manner to all subsequent layers. This process enables the extraction of information at varying coarseness levels. At the core of CondenseNeXt are several dense blocks. Additionally, it incorporates depthwise separable convolution and pooling layers between these blocks to alter feature-map dimensions accordingly. This strategy facilitates more effective extrapolation of features at diverse resolutions from an input image. Subsequently, the obtained information is then fused to mitigate the vanishing-gradient issue, resulting in efficient inference and reduced size of the trained weights because of a reduction in both the number of parameters and floating-point operations per second, as seen in [10].

3.2. Neck

In modern object detectors, the neck module plays an important role as a feature fusion hub. It collects feature maps extracted at different depths within the backbone, often using pyramid networks like feature pyramid networks (FPNs) [23] and path-aggregation networks (PANs) [24]. NeXtDet [9], for instance, leverages a combined PAN-FPN architecture. While FPNs generate feature maps of various sizes to capture diverse information, merging them can be challenging due to size discrepancies. To bridge this gap, a PAN is integrated with an FPN and upsampled using nearest-neighbor interpolation. This allows bottom-up features, rich in positioning information, to connect with top-down features, strong in semantic understanding. This fusion, visualized in Figure 4, ultimately enhances the network's performance.

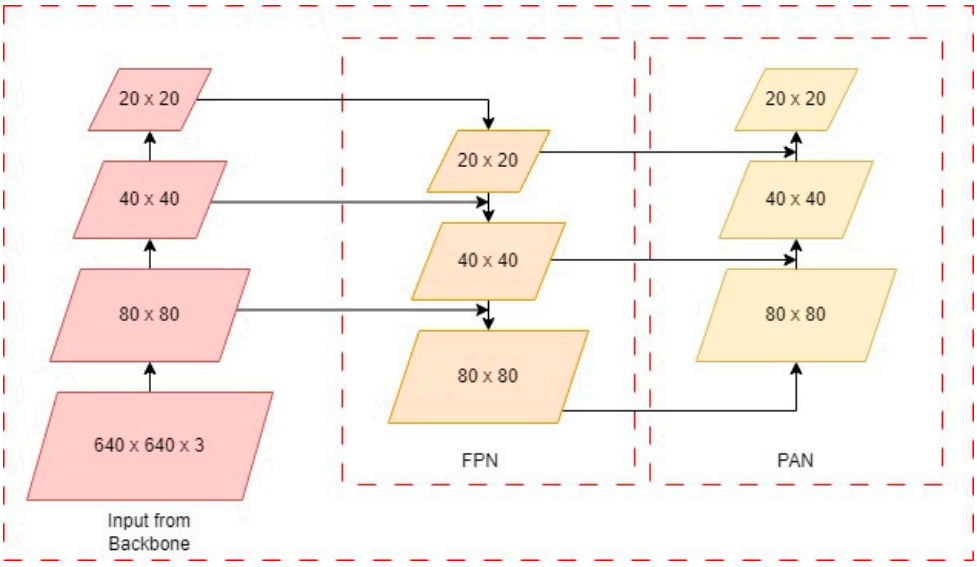

Figure 4. A visual depiction of top-down and bottom-up implementation of FPN and PAN networks.

Spatial pyramid pooling (SPP) [22] presents an innovative max-pooling technique designed to enhance the accuracy of the CNNs. It achieves this by pooling the responses of each filter within individual local spatial bins, preserving the spatial information. This concept draws inspiration from the well-known bag-of-words approach [31] in computer

vision. The strategy employs three distinct sizes of max-pooling operations to discern analogous feature maps, regardless of the diverse resolutions of input feature patterns.

After applying max pooling, the resulting information is flattened and merged before being fed into the fully connected layer. This final layer delivers an output of fixed size, independent of the initial input dimensions, as shown in Figure 5. Notably, it is the fully connected layer, not the convolution layer, that restricts the final output size of CNNs. This integration typically occurs in the later stages of feature fusion. For a visual representation, Figure 5 showcases the SPP block employed within the neck of the proposed NeXtFusion architecture.

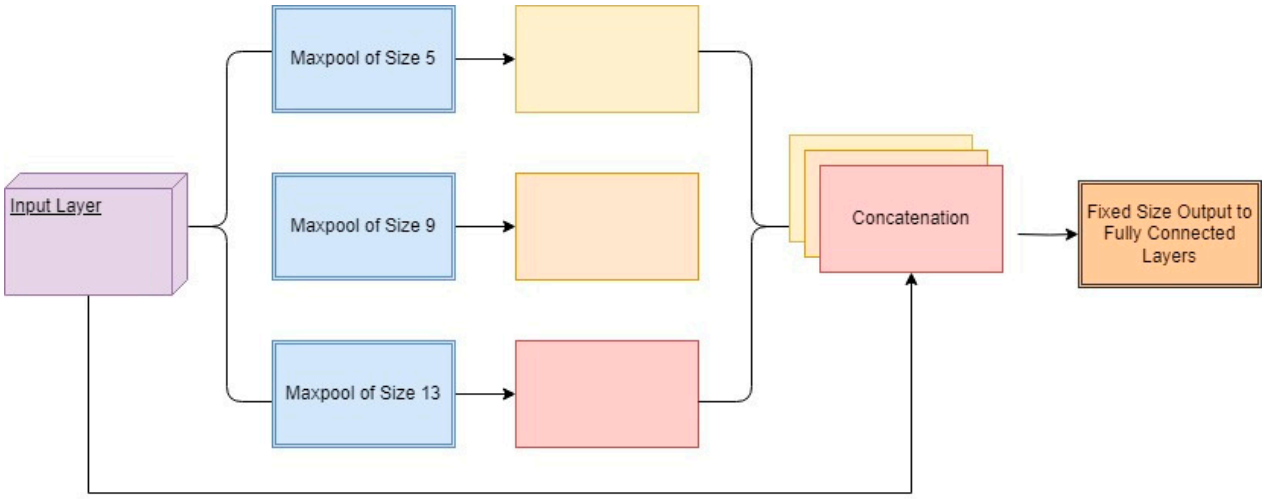

Figure 5. A visual depiction of the spatial pyramid pooling operation.

3.3. Head

The final stage of an object detector, the head module, takes center stage in defining bounding boxes and generating detailed detection results like object class, confidence score, location, and size. To achieve this, modern detectors often employ multiple head modules working together. These modules share features extracted earlier in the network and specialize in accurately identifying objects and predicting their confidence scores. In the NextDet architecture [9], three distinct heads, each equipped with a spatial attention module (SAM), tackle this crucial task. The SAM module, originally introduced in the convolutional block attention module (CBAM) [32], plays a key role in feature aggregation. It accomplishes this by creating a spatial attention map that highlights critical areas within the image. This map is generated by analyzing the relationships between different features using both max-pooling and average-pooling operations along the channel axis.

3.4. Bounding-Box Regression

The object detection task can be streamlined by dividing it into two simpler subtasks: identifying objects and pinpointing their locations. Finding these objects relies on a technique called bounding-box regression (BBR). This method essentially draws a rectangular box around the predicted object location within the image, maximizing the overlap with the actual object. The extent of this overlap is measured by mean squared error (MSE) or the intersection-over-union (IoU) losses. This popular metric evaluates the similarities and differences between two arbitrary shapes. Mathematically, it represents the ratio between the area shared by the predicted bounding box (denoted as A) and the ground-truth bounding box (B) as follows:

$$\mathrm{IoU}(\mathrm{A}, \mathrm{B}) = \frac{\mathrm{A} \cap \mathrm{B}}{\mathrm{A} \cup \mathrm{B}} \tag{1}$$

While IoU loss is widely used in BBR, it encounters a challenge when the predicted bounding box and the actual object's ground-truth box do not intersect, i.e., when IoU of A and B equals zero, it fails to provide an overlap ratio. To overcome this limitation, the NextDet object detector has integrated a more robust approach called generalized IoU (GIoU) [33]. GIoU addresses this issue by actively encouraging a greater overlap between the predicted and ground-truth boxes, effectively steering the prediction closer toward the target. Mathematically, GIoU loss ($\mathcal{L}_{\mathrm{GIoU}}$) is expressed as follows:

$$\mathcal{L}_{\mathrm{GIoU}} = 1 - \mathrm{IoU} + \frac{|\mathrm{C} - \mathrm{A} \cup \mathrm{B}|}{|\mathrm{C}|} \tag{2}$$

Here, C represents the smallest enclosing box that incorporates both the predicted bounding box (A) and the ground-truth box (B). As defined in [33], experiments reveal that GIoU loss delivers superior performance compared to both mean squared error (MSE) and standard IoU losses. Notably, it also demonstrates effectiveness in tackling vanishing gradients when the predicted and ground-truth boxes fail to intersect. Figures 6 and 7 provide a visual overview of the difference between IoU and GIoU.

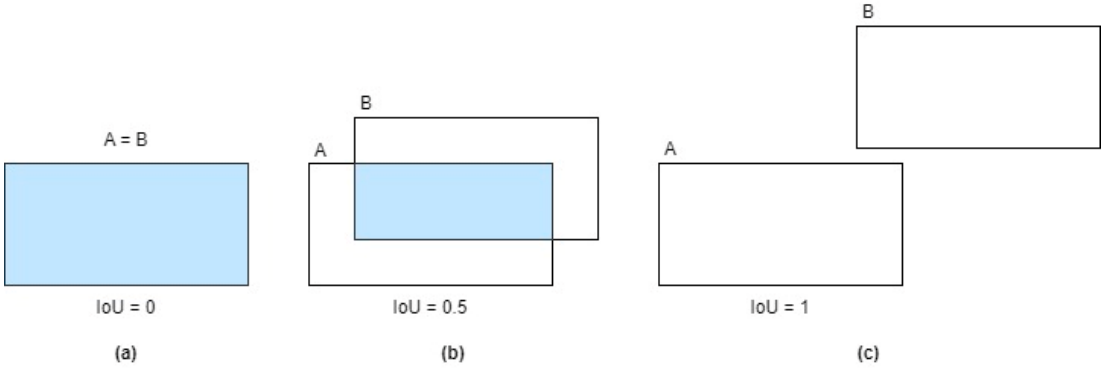

Figure 6. Three sets of examples (**a**–**c**) where (**a**) represents a perfect overlap between the predicted bounding box (A) and the ground-truth box (B), (**b**) represents a partial overlap resulting in 0.5 IoU losses, and (**c**) represents the disjoint problem of IoU when A and B do not overlap.

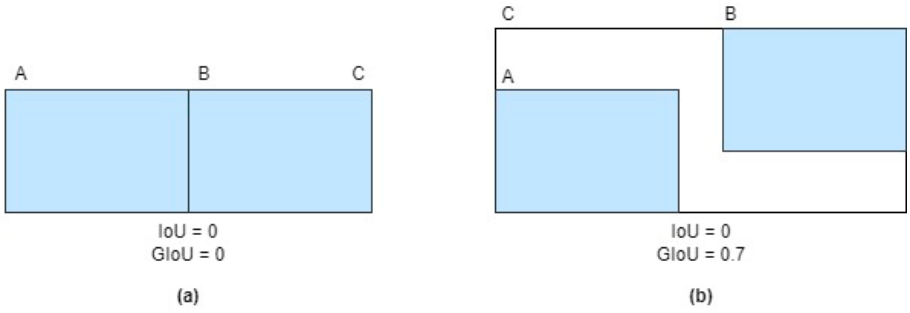

Figure 7. Two sets of examples (**a**,**b**) where (**a**) represents a perfect overlap between the predicted bounding box (A) and the ground-truth bounding box (B) and (**b**) represents a non-overlapping case of A and B, which solves the IoU's disjoint problem by introducing a third smallest enclosing bounding box called C that encompasses both A and B bounding boxes.

3.5. Extracting Radar Features

The proposed multi-modal object detection network in this paper utilizes features of each object within the image to predict all other properties associated with the object. To maximize the utilization of radar point-cloud data associated with the object in this context, it is essential to initially establish a connection between radar detections and their corresponding objects of interest detected within the image.

An autonomous vehicle's movement can be visualized using the right-handed coordinate system as it travels in the forward direction. A commonly employed method to determine the right-hand rule involves extending the index finger along the positive x-direction, bending the middle finger (and/or ring and pinky fingers) inward to indicate the positive y-direction, and raising the thumb to represent the positive z-direction, as denoted by Figure 8. In this configuration, the x-axis denotes the direction of motion, the y-axis runs parallel to the front axle of the vehicle, acting as the reference point, and the z-axis runs perpendicular to the x- and y-axes and points out through the roof of the vehicle.

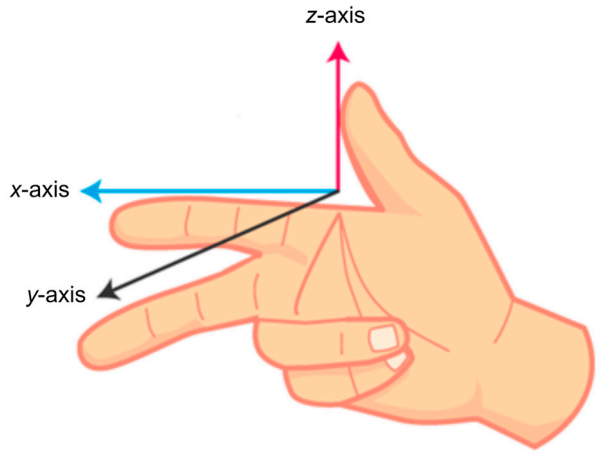

Figure 8. Right-handed coordinate system.

For handling and analyzing point-cloud data representations, i.e., data obtained from radar sensors, a polar coordinate system is being used. In this type of system, a point in space is represented by a distance (d) from a reference point (origin) and an azimuth angle ($\propto$) measured from a reference direction (usually the positive x-axis). This system is particularly suitable for representing spatial information in scenarios where the distance and angle of objects from a reference point are essential, as is often the case with radar measurements and point-cloud data. By determining the distance (r) of an object from point A and its azimuth ($\propto$) from the radar, one can make an approximation of the location of the object of interest (A) within the global coordinate system [34].

Within the global coordinate system of x, y, and z, each radar detection is expressed as a 3D point relative to the sensor's position, thus characterizing it as $(x,\ y,\ z,\ v_x,\ v_y)$, where x, y, and z denote the position of the point in a 3D space, and v_x and v_y signify the stated radial speed of the object along the x and y axes. For each scenario, three sequential radar point-cloud sweeps are combined, with a 0.25 s interval between each. Each camera within the nuScenes dataset is pre-calibrated, featuring both intrinsic and extrinsic parameters.

The intrinsic parameter is a 3×3 matrix that defines the internal characteristics of the camera, including focal length, principal point, and distortion coefficients, which are typically acquired through specialized calibration procedures that involve checkerboard patterns or similar techniques and can be defined as follows:

$$Intrinsic_{param} = I_p = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{3}$$

Here, f_x and f_y represent the focal lengths in the x and y planes, respectively, and c_x and c_y represent the offset points of the camera in the x and y planes, respectively. On the other hand, the extrinsic parameter is a combination of a rotation matrix and a translation vector that define the camera's position and orientation with respect to the vehicle position and, thus, plays a vital role in the projection of radar detections from global coordinates

onto the camera's image plane. Extrinsic parameters for a camera in the nuScenes dataset can be defined as follows:

$$Extrinsic_{param} = E_p = \begin{bmatrix} r_{xx} & r_{xy} & r_{xz} & t_x \\ r_{yx} & r_{yy} & r_{yz} & t_y \\ r_{zx} & r_{zy} & r_{zz} & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (4)$$

Here, the nine elements, r_{xx} to r_{zz}, represent the 3×3 rotation matrix that describes the camera's orientation relative to the world coordinate system, and the three elements, t_x to t_z, represent the translation vector that describes the camera's position in 3D space relative to the world coordinate system. These calibration data are provided by the nuScenes dataset [35] and are utilized along with the camera-radar dataset for experiments outlined in Section 4 of this paper. The radar detections can, therefore, be associated with their corresponding representations obtained from the camera sensor. Following this mapping process, detections located outside the image are discarded. The projection of radar detections from global coordinates onto the camera's image plane can be defined as follows:

$$P_{camera} = I_p \times E_p \times P_{world} \quad (5)$$

Here, P_{camera} is a 3×1 vector representing a 3D coordinate system of the camera, i.e., $\begin{bmatrix} x & y & z \end{bmatrix}^T$, and P_{world} is a 4×1 vector representing a 3D point in the global coordinate system, i.e., $\begin{bmatrix} X & Y & Z & 1 \end{bmatrix}^T$.

3.6. Assosciating Radar Data to the Image Plane

The proposed NeXtFusion network utilizes a modified frustum generation mechanism to associate image data with radar data, similar to CenterFusion's approach [36]. This technique leverages the object's two-dimensional bounding box from the camera sensor along with estimations of the object's three-dimensional size, depth, and orientation from the radar sensor. By doing so, a tightly defined region of interest (RoI) is constructed around the object of interest called frustum. This frustum then facilitates the filtering of radar detections. Only radar detections located within the frustum are then considered for association (concatenation) with the camera-detected object.

Instead of utilizing only one radar detection for each object proposal individually, as described in [36], the proposed NeXtFusion network utilizes a modified mechanism that utilizes an entire cluster of radar detections that fall within the object's designated RoI. This allows the network to make use of collective information within the cluster, resulting in a more robust camera-radar data association because it was observed that the entire radar detection cluster, encompassing its shape, size, and orientation, holds valuable information resulting in an improvement in multi-modal object detection compared to the approach described in [36] which focuses solely on individual detections that only provide information about their specific location and velocity.

Figure 9 provides a visual representation of the architecture of the proposed NeXtFusion multi-modal (sensor fusion) 3D object detection network. The design of the proposed network incorporates several modifications to the baseline single-modal (camera-based) NextDet [9] architecture. The proposed architecture comprises two CondenseNeXt [25] CNNs integrated into the backbone module to extract feature map representations from data acquired through camera and radar sensors. These feature maps are subsequently transmitted to the neck for feature fusion, facilitated by connections indicated by dashed lines. Furthermore, the design of the network incorporates a strategy called early fusion, where feature maps from cameras and radars are concatenated at an initial stage. This merges information from both sensors (camera for visual details and radar for object distance and presence) to create a richer feature representation for object detection. Thus, by combining these features early on, the network can learn a more robust representation of the environment for object detection.

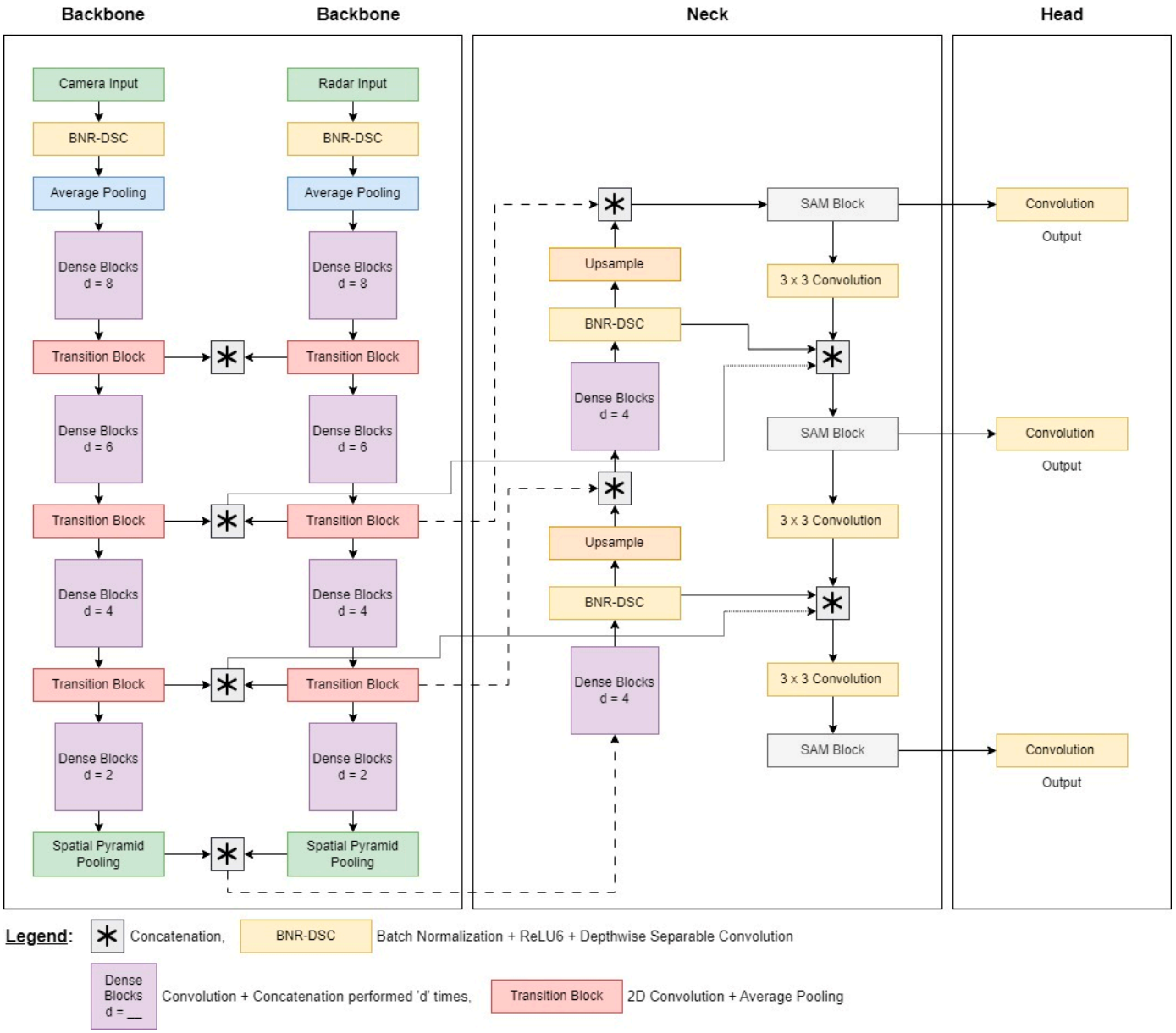

Figure 9. Architecture of the proposed NeXtFusion 3D object detection network.

Inspired by the work of [37], the design of NeXtFusion incorporates these connections in the network, which aims to improve gradient backpropagation, alleviate gradient fading, and minimize the loss of feature information, especially in scenarios involving small objects in adverse weather conditions, resulting in an improved and robust performance of the proposed object detection network.

4. Experiments and Results

A modern object detector is usually designed to identify the location and type of object present in each input image. Typically, such a detector undergoes training on a dataset consisting of labeled images, referred to as ground-truth values. In this section, the proposed network, NeXtFusion, is evaluated on the nuScenes dataset as part of the research presented within this paper. A comparative analysis is conducted on the proposed network and compared to the other existing object detection neural networks in Section 4.4. Additionally, samples from the nuScenes dataset are visualized in Section 4.5 to better understand the object detection and tracking performance of the proposed network.

4.1. Dataset

Extensive experiments are performed on the nuScenes dataset [35], a multi-modal dataset for autonomous driving, which provides challenging urban driving scenarios using the full suite of sensors from a real autonomous vehicle. It offers annotated images, bounding boxes, and point-cloud radar data suitable for object detection, tracking, and forecasting tasks related to autonomous driving, everyday objects, and humans using cameras and radar sensors. Figure 10 provides a sample from the nuScenes dataset.

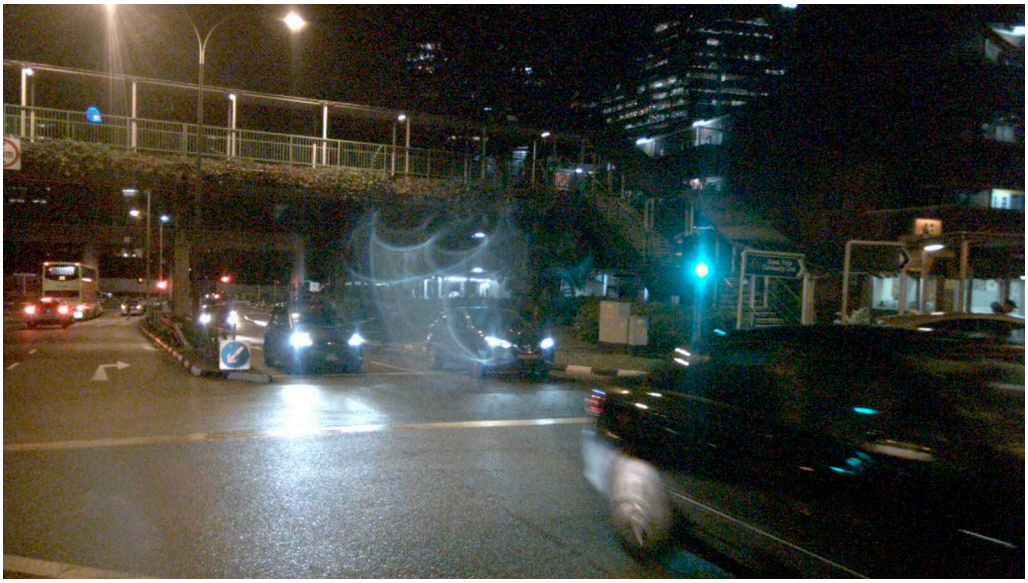

Figure 10. An example of an occluded image obtained from the nuScenes dataset, captured by a front camera positioned at the vehicle's top-front location.

4.2. Evaluation Metrics

This paper establishes two main criteria for an object detector to be considered successful in identifying target objects using the proposed multi-modal detection approach. Notably, a GIoU threshold of 0.5 is consistently applied across all models and datasets examined, employing a grid-based search approach to ensure consistency. This metric extends IoU by considering the minimum bounding box that can enclose both the predicted and ground-truth boxes, as explained in Section 3.4 within this paper. This technique penalizes predictions that are far away from the ground truth, even if they have some overlap.

Evaluating the performance of object detectors requires a variety of metrics [38]. A popular choice for both assessment and comparison is the mean average precision (mAP). This metric represents the average performance across all object categories, calculated by taking the mean of the average precision (AP) for each class. AP itself is derived from the area under the precision-recall (PR) curve. Within this curve, precision (P) reflects a model's ability to correctly identify true objects, indicating the percentage of predictions that are positive. Conversely, recall (R) assesses a model's ability to find all actual positive instances present in the ground-truth data. These metrics can be expressed mathematically as follows:

$$P = \frac{TP}{TP + FP} \tag{6}$$

$$R = \frac{TP}{TP + FN} \tag{7}$$

$$AP = \frac{\Sigma P}{number\ of\ objects} \tag{8}$$

$$mAP = \frac{\Sigma AP}{number\ of\ classes} \tag{9}$$

Here, the variables represent the following:

- True positive (TP) represents correctly identified objects where the predicted bounding box overlaps with the ground-truth box.
- False positive (FP) denotes erroneous detections where the predicted box has some overlap with the ground truth but falls, indicating an incorrect identification.
- False negative (FN) occurs when the model fails to detect a valid object present in the scene, resulting in a missed target.

Additionally, training parameters for experimental analyses are outline in Table 1 and average errors in different aspects of 3D object detection are evaluated on the nuScenes dataset in Table 2. The benchmarking metrics utilized for these analyses are as follows:

- Mean average translation error (mATE): This metric represents the average Euclidean distance (in meters) between the predicted and ground-truth 3D translation (x, y, z) of an object across all objects and scenes in the dataset.
- Mean average scale error (mASE): This metric represents the average absolute difference between the predicted and ground-truth scale (measured as the product of length, width, and height) of an object across all objects and scenes.
- Mean average orientation error (mAOE): This metric represents the average minimum yaw angle difference (in radians) between the predicted and ground-truth orientation of an object across all objects and scenes. It is typically calculated on a per-class basis due to variations in object types.
- Mean average velocity error (mAVE): This metric represents the average absolute difference between the predicted and ground-truth velocity (magnitude in m/s) of an object across all objects and scenes. It is usually calculated on a per-class basis.
- Mean average attribute error (mAAE): This metric represents the average absolute difference between the predicted and ground-truth attributes (depending on the specific attribute considered) of an object across all objects and scenes. It is typically calculated on a per-class basis and depends on the specific set of attributes considered for the object category.

These metrics are commonly used in the nuScenes dataset to evaluate the performance of object detection and tracking algorithms, particularly when dealing with 3D object detection and perception tasks. They provide insights into the overall accuracy and robustness of the algorithms regarding various aspects like their ability to localize objects correctly, estimate their size and orientation accurately, track their movement, and even predict certain attributes depending on the task.

4.3. Training Infrastructure

Experiments outlined in this paper utilize the Carbonate supercomputer's GPU partition, a powerful research cluster specialized for deep-learning tasks. Located at Indiana University, this resource offers 24 GPU-accelerated nodes, each equipped with cutting-edge hardware as follows:

- NVIDIA Tesla V100 GPU: Delivers high-performance graphics processing for efficient training.
- Intel Xenon Gold 6248 20-core CPU: Provides robust central processing power for computations.
- 1.92 TB solid-state drive: Ensures fast data storage and retrieval.
- 768 GB of RAM: Supports large model training and data handling.

This research was supported in part by Shared University Research grants from IBM Inc. and Lilly Endowment Inc. and by Indiana University's Pervasive Technology Institute [39]. Additionally, the experiments employed the following:

- PyTorch 1.12.1: A deep-learning framework for model development and training.

- Python 3.7.9: The general-purpose programming language used for the research.
- CUDA 11.3: Enables efficient utilization of the NVIDIA GPUs for computations.

4.4. Experiment Results

To assess the performance of the proposed multi-modal 3D object detection network and understand the benefit of multi-modal over single-modal object detection networks, NeXtFusion has been compared to the camera-based 3D object detection networks, orthographic feature transform (OFT) [40] and monocular 3D object detection (MonoDIS) [41], as well as InfoFocus [42], which is a LiDAR-based 3D object detection neural network, and benchmarked on the nuScenes dataset.

The effectiveness of any object detection network relies heavily on its training process. Table 1 presents the training parameters used for both the existing and the proposed object detection networks analyzed in this study. As is evident from this table, all network models employ the Adam optimizer, a learning rate of 0.0001, batch size of 64, weight decay of 0.0005, momentum of 0.85, and trained for 400 epochs.

Table 1. Training prameters for experimental analysis on the nuScenes training dataset.

Object Detection Network	Training Dataset	Optimizer	Learning Rate	Batch Size	Weight Decay	Momentum	# of Epochs
OFT	Y	Adam	0.0001	64	0.0005	0.85	400
MonoDIS	Y	Adam	0.0001	64	0.0005	0.85	400
InfoFocus	Y	Adam	0.0001	64	0.0005	0.85	400
NeXtFusion	Y	Adam	0.0001	64	0.0005	0.85	400

Table 2 presents a comparison of several 3D object detection methods on the nuScenes validation dataset, indicating performance metrics such as mAP, mATE, mASE, mAOE, mAVE, and mAAE described in Section 4.2 of this paper. The models compared are 3D object detection networks that utilize data primarily from camera or lidar sensors such as OFT, MonoDIS, and InfoFocus. As is evident in this table, the proposed NeXtFusion network achieves a remarkable performance, securing the highest mAP score among all compared 3D object detection networks. This metric signifies the overall accuracy of the proposed multi-modal object detection model across different confidence levels.

Table 2. Comparison of 3D Object Detection Methods on the nuScenes Validation Dataset (↑ indicates Higher = Better, ↓ indicates Lower = Better, and Y indicates the use of nuScenes validation dataset).

Object Detection Network	Validation Dataset	mAP ↑	mATE ↓	mASE ↓	mAOE ↓	mAVE ↓	mAAE ↓
OFT	Y	0.122	0.819	0.359	0.848	1.727	0.479
MonoDIS	Y	0.301	0.737	0.266	0.544	1.531	0.140
InfoFocus	Y	0.396	0.361	0.263	1.130	0.617	0.394
NeXtFusion	Y	0.473	0.449	0.261	0.534	0.538	0.139

Compared to its single sensor-based 3D object detection competitors, the proposed NeXtFusion network exhibits significant improvements in mAP. Notably, it outperforms MonoDIS by 9.5% and surpasses OFT by a margin of 35.1%. While InfoFocus exhibits a very strong mAP score, its reliance solely on LiDAR data limits its capabilities compared to NeXtFusion's fusion of multiple sensor modalities. This multi-sensor (multi-modal) approach enables NeXtFusion to achieve a significant improvement in velocity error compared to camera and lidar-based methods, demonstrating the added value of utilizing

diverse sensor information for a more comprehensive understanding of the environment in adverse external environmental conditions.

4.5. Ablation Studies

Table 3 outlines ablation studies conducted on the nuScenes validation dataset to evaluate the effectiveness of the proposed multi-modal fusion network in this paper. This table outlines the single-modal object detection performance of two networks: MonoDIS and NeXtDet. It serves as a baseline for understanding how NeXtFusion, which builds upon the efficient NeXtDet architecture, makes use of additional sensor (radar) data to improve object detection performance. While both networks achieve reasonable mean average precision (mAP) scores, indicating their ability to identify objects, NeXtDet demonstrates a 16.61% improvement over MonoDIS. This suggests that NeXtDet's architecture is better suited for extracting relevant features obtained from the sensor data, resulting in an increased accuracy in object detection (mAP). However, the differences in other metrics are very subtle. While MonoDIS exhibits slightly lower errors in terms of mATE, mASE, and mAOE, these differences are negligible. Conversely, NeXtDet shows a slight improvement in mAVE and mAAE.

Table 3. Results of the ablation studies on nuScenes validation dataset (C: camera, R: radar, NM: naïve method, and FAM: frustum association method).

Network	C	R	NM	FAM	mAP ↑	mATE ↓	mASE ↓	mAOE ↓	mAVE ↓	mAAE ↓
MonoDIS	✓	-	-	-	0.301	0.738	0.269	0.544	1.532	0.141
NeXtDet	✓	-	-	-	0.351	0.635	0.271	0.550	1.346	0.142
NeXtFusion	✓	✓	✓	-	0.427	0.499	0.268	0.541	0.846	0.140
NeXtFusion	✓	✓	-	✓	0.474	0.448	0.262	0.534	0.536	0.138

The second half of the ablation study focuses on the proposed NeXtFusion network and the impact of two crucial methods responsible for associating radar detections with objects in the image plane, especially for multi-modal networks. The first is the naïve method (NM), where each radar detection point is projected directly onto the image plane using the sensor calibration information. Therefore, if the projected radar point falls within the two-dimensional bounding box of the detected object inside an image, then it is associated with that object. NM is compared to the modified frustum association method (FAM) outlined in Section 3.5 of this paper and utilized in the design of the proposed NeXtFusion network. The results of this study demonstrate the potential benefits of sensor fusion. Compared to the camera-only results, NeXtFusion achieves significant improvements in several key metrics when utilizing FAM, demonstrating a substantial increase in the network's ability to correctly identify objects. Additionally, there are considerable reductions in errors related to object location, scale, and orientation, as observed in Table 3.

4.6. Visualization of Samples

The nuScenes dataset, with its diverse collection of 1000 driving scenes, captures various scenarios that provide a platform to evaluate, benchmark, and compare 3D object detection algorithms. Each scene in this dataset is roughly 20 s long and provides information, including camera images, radar data, and meticulously labeled objects. Although LiDAR data are also represented as a point-cloud representation, similar to radar data in this dataset, LiDAR and radar point-cloud data are fundamentally different despite both appearing as point clouds. LiDAR provides highly detailed 3D representations while radar excels at long-range detection, and therefore, the work presented within this paper focuses on camera-radar sensor fusion. Hence, LiDAR data are not utilized for experiments outlined in this section.

This section delves into visualizing the proposed network's validation performance in terms of 3D object detection on the nuScenes dataset. In this dataset, the scenes are annotated every half a second (i.e., at 2 Hz). A sample is defined as an annotated keyframe of a scene at a specific timestamp, where the timestamps of data from all sensors closely align with the sample's timestamp. For illustration, let us examine the first annotated sample in a scene described as 'Night, pedestrians on sidewalk, pedestrians crossing crosswalk, scooter, with difficult lighting conditions' by the annotators of the dataset. This scene was captured in Holland Village, Singapore, using a comprehensive sensor suite mounted on a vehicle. Each snapshot of a scene references a collection of data from these sensors, accessible through a data key in the dataset. These sensors, which are mounted on the vehicle, include the following:

- One LIDAR (light detection and ranging) sensor:
- LIDAR_TOP
- Five RADAR (radio detection and ranging) sensors:
- RADAR_FRONT
- RADAR_FRONT_LEFT
- RADAR_FRONT_RIGHT
- RADAR_BACK_LEFT
- RADAR_BACK_RIGHT
- Six camera sensors:
- CAM_FRONT
- CAM_FRONT_LEFT
- CAM_FRONT_RIGHT
- CAM_BACK
- CAM_BACK_LEFT
- CAM_BACK_RIGHT

Figure 11 provides a visual representation of 3D object detection performed by the proposed NeXtFusion network on a scene with occlusion and poor lighting conditions, and Figure 12 provides an example of multi-modal 3D object detection involving multiple objects at an intersection, such as humans (blue bounding box), cars (yellow bounding box), and trucks (red bounding box). It also demonstrates the object detection performance of distant, tiny objects in the shadow of the building.

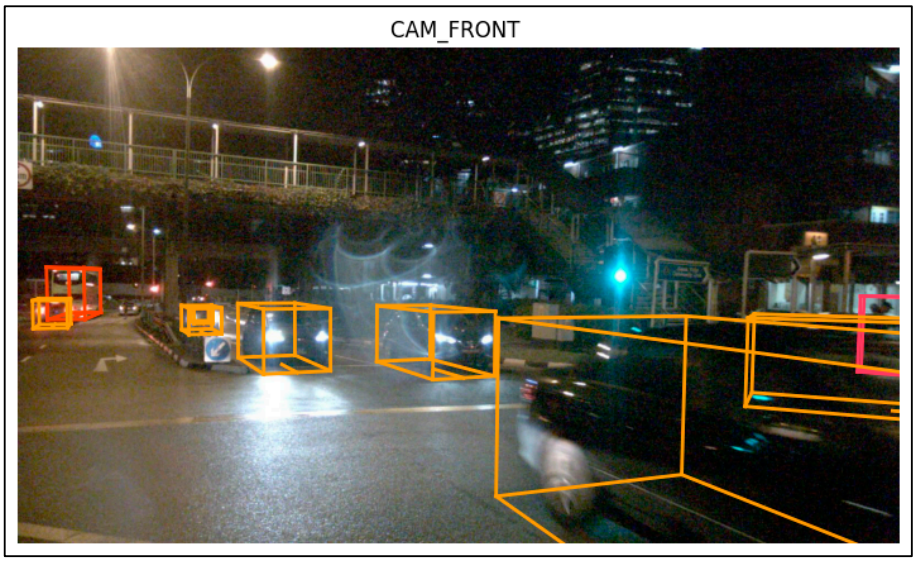

Figure 11. An example of Camera-Radar fusion-based 3D object detection involving occlusion at night from the nuScenes dataset using the proposed NeXtFusion network. Different colors of the bounding boxes indicate different objects detected.

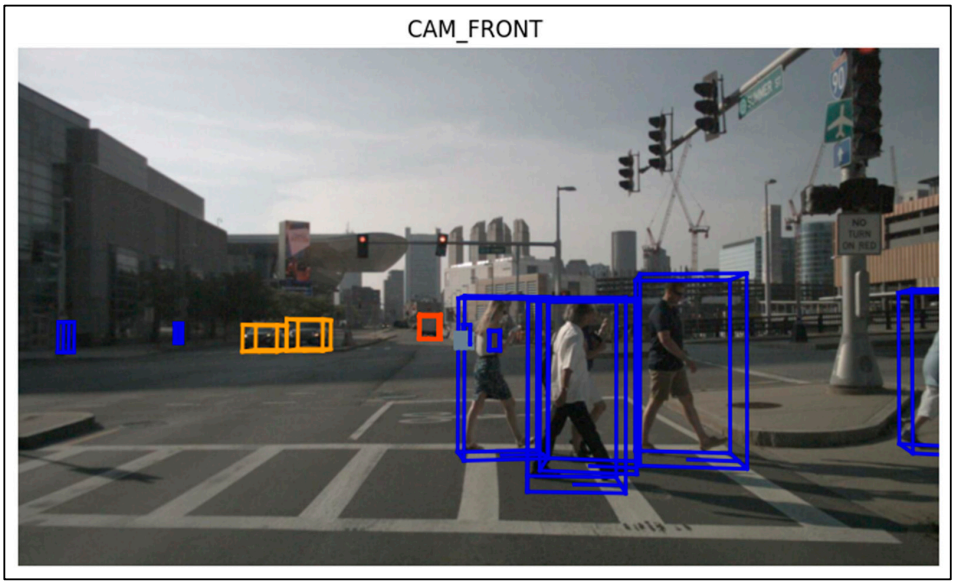

Figure 12. An example of Camera-Radar fusion-based 3D object detection at an intersection using the proposed NeXtFusion network. Here, blue bounding boxes, denote people, yellow bounding boxes denote cars and red bounding boxes denote bus.

Figure 13 provides an example that plots radar point-cloud data for the same image from Figure 12. Unlike LiDAR, which excels in dense, close-range measurements, radar boasts a significantly larger operational range. While this extended reach comes at the expense of point density, it enables the detection of distant objects that might be invisible to LiDAR. Consider a scenario where a car is approaching a sharp bend on a highway. While LiDAR can meticulously map the immediate surroundings, radar's wider net can detect vehicles or obstacles further down the road, providing crucial information for safe navigation. This complementary perspective offered by radar, despite its sparser data, paints a more comprehensive picture of the environment, enhancing perception and decision-making capabilities in scenarios demanding long-range awareness.

Figure 13. An illustration of radar point-cloud data of the environment from the nuScenes dataset generated by NeXtFusion. Dots represent objects, with darker shades indicating closer proximity to the sensor.

Figure 14 illustrates a point-cloud plot generated by combining radar data from five sweeps. This data is obtained from the nuScenes dataset utilized for experiments outlined in this paper. As seen in this plot, radar detection lines visually align with their respective bounding boxes of two vehicles, providing an estimate of the detected objects. Figure 15 shifts the evaluation and visualization focus to object tracking, for which a tracking pipeline was developed for evaluation purposes. By providing more reliable starting points in each frame, the proposed network analyzes data about objects, which, in other words, are individual entities within the environment that an autonomous vehicle (AV) needs to detect, track, and potentially interact with based on the conditions. Examples include specific vehicles, pedestrians, traffic signs, or other relevant objects. Understanding these instances and their associated metadata is crucial for safe and efficient navigation.

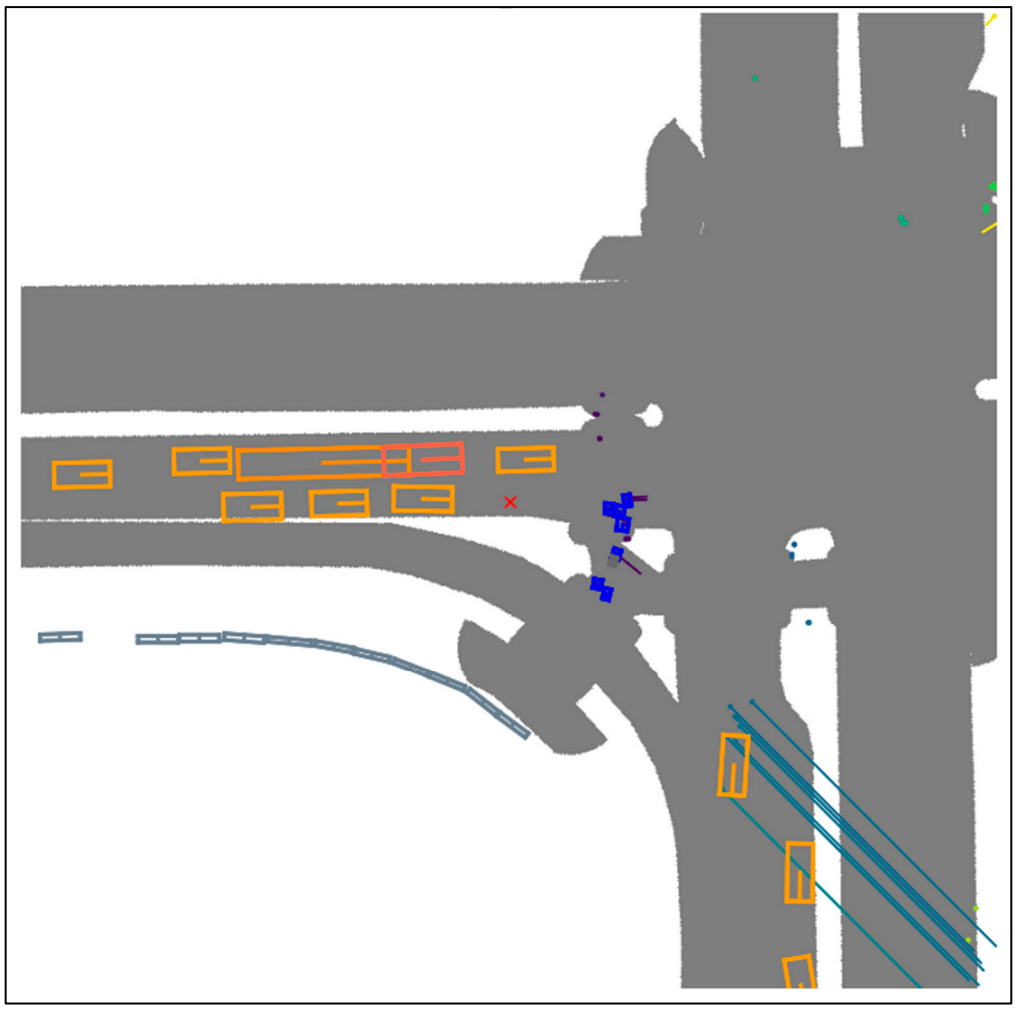

Figure 14. An example of very confident radar returns from two vehicles, captured and combined using five radar sweeps. Light blue lines indicate radar detection, indicating its length and orientation. Red cross indicates vehicle's ego-centric position. Here, blue bounding boxes, denote people, yellow bounding boxes denote cars and red bounding boxes denote bus.

Consider a hypothetical scenario where the camera sensor is affected by lens glare from the sun rays directly hitting the camera sensor, resulting in an occlusion in the field of vision. Here, an AV encounters pedestrians crossing an intersection on the road. Examining the instance metadata associated with this object unveils valuable information beyond its simple presence, even in adverse conditions such as lens glare. These rich metadata are generated by fusing information from radar and cameras in NeXtFusion, which plays a critical role in AV perception and decision making. By tracking and analyzing instance metadata, the AV can accomplish the following:

- Continuously monitor the movement of objects in its environment.
- Classify and differentiate between different types of objects and understand their potential intentions even under unfavorable conditions.
- Make informed decisions by planning safe maneuvers based on perceived information about the environment.

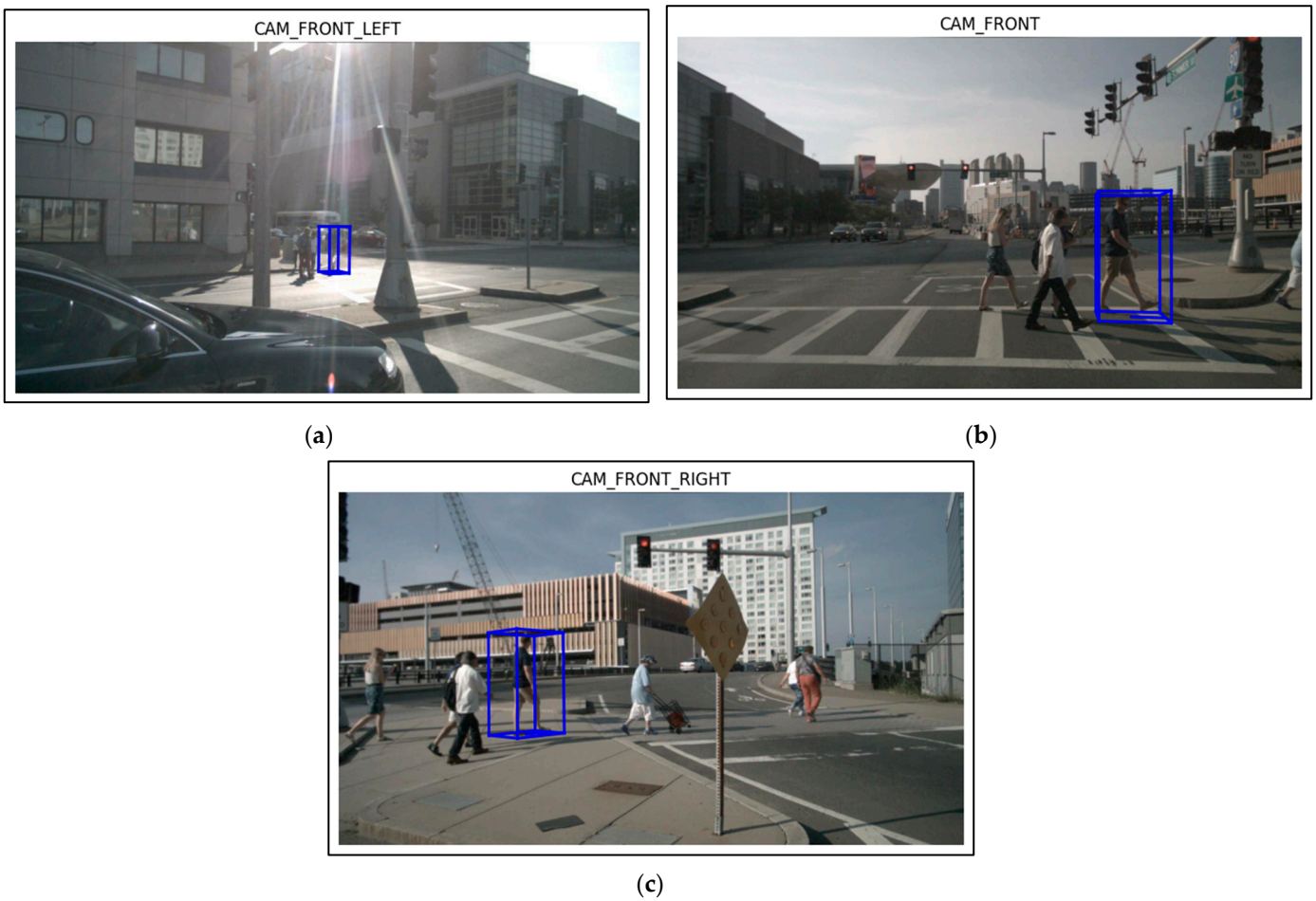

Figure 15. A static representation to analyze how the proposed network performs multi-modal 3D object detection and tracking as part of the validation tests. nuScenes provides data for a scene captured from camera and radar sensors mounted on the AV as follows: (**a**) Front left camera identifies and starts tracking the object. In this example, only a single pedestrian is being tracked for analysis purposes. (**b**) Front center camera continues to track the identified object while it is in its field of view. (**c**) Front right camera continues to track the identified object until it leaves its field of view.

While Figure 15 serves as a static representation to understand how the proposed network performs 3D object detection and tracking, the AV perception is always dynamic in nature and performs object detection and tracking for each frame. The work presented within this paper focuses on processing individual camera images and radar scans simultaneously. Also, metadata continuously updates as sensors gather new information, allowing the AV to adapt its understanding of the surrounding world in real time. This dynamic interplay between sensor data, object detection and tracking, and rich metadata forms the foundation for safe and intelligent navigation for autonomous driving.

5. Conclusions

This paper introduces NeXtFusion, a novel deep camera-radar fusion network designed to enhance the perception capabilities of autonomous vehicles (AVs). By effectively combining the strengths of camera and radar data, NeXtFusion overcomes the limitations

of single-sensor networks, particularly in challenging weather and lighting conditions. Utilizing an attention module, NeXtFusion extracts crucial features from both modalities, leading to improved object detection accuracy and tracking.

Extensive evaluations on the nuScenes dataset demonstrate NeXtFusion's superior performance, achieving a significant mAP score improvement compared to existing methods. Additionally, strong performance in other metrics like mATE and mAOE further highlights its overall effectiveness in 3D object detection. Visualizations of real-world data processed by NeXtFusion showcase its ability to handle diverse scenarios. By leveraging the complementary information from cameras and radars, NeXtFusion offers a robust solution for navigating complex environments and ensuring reliable operation under various conditions.

However, recent research explores the potential of utilizing additional sources of information to enhance an AV's awareness of its perceived surroundings, including but not limited to roadside sensors, such as surveillance cameras and radars, along with unmanned aerial vehicles (UAVs). These technologies contribute to the topic of digitalization of traffic scenes, providing more comprehensive information about the surrounding environment to the AVs. By establishing communication between AVs and this intelligent infrastructure, researchers further envision a future where object detection is not just reliant only on onboard sensors but off-vehicle sensors as well.

The research work and findings presented in this paper, particularly the ability to combine and extract meaningful and complementary information from camera and radar sensors, could potentially contribute to the development of such infrastructure-vehicle cooperation systems. Exploring this exciting potential application is a promising avenue for future research.

Author Contributions: Conceptualization, P.K. and M.E.-S.; methodology, P.K.; software, P.K.; validation, P.K. and M.E.-S.; writing—original draft preparation, P.K.; writing—review and editing, P.K. and M.E.-S.; supervision, M.E.-S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The openly available public dataset nuScenes has been utilized in this study. It is cited as [35] of this paper.

Acknowledgments: The authors would like to acknowledge the Indiana University Pervasive Technology Institute for providing supercomputing and storage resources as well as the Internet of Things (IoT) Collaboratory at the Purdue School of Engineering and Technology at Indiana University Purdue University at Indianapolis that have contributed to the research results reported within this paper.

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

References

1. Robson, K. Full Self-Driving Cars Might Not Be with Us until 2035, Experts Predict. Available online: https://www.verdict.co.uk/fully-self-driving-cars-unlikely-before-2035-experts-predict/ (accessed on 21 December 2023).
2. Tang, Y.; He, H.; Wang, Y.; Mao, Z.; Wang, H. Multi-Modality 3D Object Detection in Autonomous Driving: A Review. *Neurocomputing* **2023**, *553*, 126587. [CrossRef]
3. Qian, R.; Lai, X.; Li, X. 3D Object Detection for Autonomous Driving: A Survey. *Pattern Recognit.* **2022**, *130*, 108796. [CrossRef]
4. Le, H.-S.; Le, T.D.; Huynh, K.-T. A Review on 3D Object Detection for Self-Driving Cars. In Proceedings of the 2022 RIVF International Conference on Computing and Communication Technologies (RIVF), Ho Chi Minh City, Vietnam, 20–22 December 2022; pp. 398–403.
5. Alessandretti, G.; Broggi, A.; Cerri, P. Vehicle and Guard Rail Detection Using Radar and Vision Data Fusion. *IEEE Trans. Intell. Transp. Syst.* **2007**, *8*, 95–105. [CrossRef]
6. Zhou, Y.; Liu, L.; Zhao, H.; López-Benítez, M.; Yu, L.; Yue, Y. Towards Deep Radar Perception for Autonomous Driving: Datasets, Methods, and Challenges. *Sensors* **2022**, *22*, 4208. [CrossRef] [PubMed]
7. Wang, Y.; Chao, W.-L.; Garg, D.; Hariharan, B.; Campbell, M.; Weinberger, K.Q. Pseudo-LiDAR From Visual Depth Estimation: Bridging the Gap in 3D Object Detection for Autonomous Driving. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 8437–8445.

8. Ku, J.; Mozifian, M.; Lee, J.; Harakeh, A.; Waslander, S.L. 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; pp. 1–8.
9. Kalgaonkar, P.; El-Sharkawy, M. NextDet: Efficient Sparse-to-Dense Object Detection with Attentive Feature Aggregation. *Future Internet* **2022**, *14*, 355. [CrossRef]
10. Kalgaonkar, P. AI on the Edge with CondenseNeXt: An Efficient Deep Neural Network for Devices with Constrained Computational Resources. Master's Thesis, Purdue University Graduate School, Indianapolis, IN, USA, 2021.
11. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
12. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016, Proceedings, Part I 14*; Springer International Publishing: Cham, Switzerland, 2016; Volume 9905, pp. 21–37.
13. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision, Salt Lake City, UT, USA, 18–23 June 2018.
14. Law, H.; Deng, J. CornerNet: Detecting Objects as Paired Keypoints. In Proceedings of the European Conference on Computer Vision (ECCV), Seoul, Republic of Korea, 27 October–2 November 2019.
15. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. CenterNet: Keypoint Triplets for Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019.
16. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015.
17. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [CrossRef] [PubMed]
18. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 24 January 2018; pp. 2980–2988.
19. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving into High Quality Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
20. Pang, J.; Chen, K.; Shi, J.; Feng, H.; Ouyang, W.; Lin, D. Libra R-CNN: Towards Balanced Learning for Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
21. Kim, J.; Sung, J.-Y.; Park, S. Comparison of Faster-RCNN, YOLO, and SSD for Real-Time Vehicle Type Recognition. In Proceedings of the 2020 IEEE International Conference on Consumer Electronics—Asia (ICCE-Asia), Seoul, Republic of Korea, 1–3 November 2020; pp. 1–4.
22. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. In *Proceedings of the Computer Vision—ECCV 2014*; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 346–361.
23. Lin, T.-Y.; Dollár, 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.
24. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
25. Kalgaonkar, P.; El-Sharkawy, M. CondenseNeXt: An Ultra-Efficient Deep Neural Network for Embedded Systems. In Proceedings of the 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 27–30 January 2021; pp. 0524–0528.
26. Modas, A.; Sanchez-Matilla, R.; Frossard, P.; Cavallaro, A. Towards Robust Sensing for Autonomous Vehicles: An Adversarial Perspective. *IEEE Signal Process. Mag.* **2020**, *37*, 14–23. [CrossRef]
27. Hoermann, S.; Henzler, P.; Bach, M.; Dietmayer, K. Object Detection on Dynamic Occupancy Grid Maps Using Deep Learning and Automatic Label Generation. In Proceedings of the 2018 IEEE Intelligent Vehicles Symposium (IV), Changshu, China, 26–30 June 2018; pp. 826–833.
28. El Natour, G.; Bresson, G.; Trichet, R. Multi-Sensors System and Deep Learning Models for Object Tracking. *Sensors* **2023**, *23*, 7804. [CrossRef]
29. Srivastav, A.; Mandal, S. Radars for Autonomous Driving: A Review of Deep Learning Methods and Challenges. *IEEE Access* **2023**, *11*, 97147–97168. [CrossRef]
30. Huang, G.; Liu, Z.; Maaten, L.V.D.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269.
31. Zhang, Y.; Jin, R.; Zhou, Z.-H. Understanding Bag-of-Words Model: A Statistical Framework. *Int. J. Mach. Learn. Cyber.* **2010**, *1*, 43–52. [CrossRef]
32. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In *Proceedings of the Computer Vision—ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 3–19.

33. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 658–666.
34. Llinas, M.L.I.; David Hall, J. (Eds.) *Handbook of Multisensor Data Fusion: Theory and Practice*, 2nd ed.; CRC Press: Boca Raton, FL, USA, 2017; ISBN 978-1-315-21948-6.
35. Caesar, H.; Bankiti, V.; Lang, A.H.; Vora, S.; Liong, V.E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; Beijbom, O. nuScenes: A Multimodal Dataset for Autonomous Driving. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11618–11628.
36. Nabati, R.; Qi, H. CenterFusion: Center-Based Radar and Camera Fusion for 3D Object Detection. In Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2021; pp. 1526–1535.
37. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
38. Padilla, R.; Passos, W.L.; Dias, T.L.B.; Netto, S.L.; da Silva, E.A.B. A Comparative Analysis of Object Detection Metrics with a Companion Open-Source Toolkit. *Electronics* **2021**, *10*, 279. [CrossRef]
39. Stewart, C.A.; Welch, V.; Plale, B.; Fox, G.; Pierce, M.; Sterling, T. Indiana University Pervasive Technology Institute. 2017. Available online: https://scholarworks.iu.edu/dspace/items/ddb55636-7550-471d-be5f-d9df6ee82310 (accessed on 26 March 2024).
40. Roddick, T. Orthographic Feature Transform for Monocular 3D Object Detection. *arXiv* **2018**, arXiv:1811.08188.
41. Chen, X.; Kundu, K.; Zhang, Z.; Ma, H.; Fidler, S.; Urtasun, R. Monocular 3D Object Detection for Autonomous Driving. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2147–2156.
42. Wang, J.; Lan, S.; Gao, M.; Davis, L.S. InfoFocus: 3D Object Detection for Autonomous Driving with Dynamic Information Modeling. In *Proceedings of the Computer Vision—ECCV 2020*; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 405–420.