IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 16 December 2023, accepted 28 February 2024, date of publication 4 March 2024, date of current version 8 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3373310

RESEARCH ARTICLE

Semi-Automatic Annotation of 3D Radar and Camera for Smart Infrastructure-Based Perception

SHIVA AGRAWAL[1], SAVANKUMAR BHANDERI[1], AND GORDON ELGER[1,2]
[1]Institute of Innovative Mobility, Technische Hochschule Ingolstadt, 85049 Ingolstadt, Germany
[2]Fraunhofer IVI, Applied Center Connected Mobility and Infrastructure, 85051 Ingolstadt, Germany

Corresponding author: Shiva Agrawal (shiva.agrawal@thi.de)

This research work is supported by the Bavarian Ministry of Economic Affairs, Regional Development and Energy (StMWi) in the project "INFRA–Intelligent Infrastructure" and the publication is partly supported by the Open Access Publication Fund of Technische Hochschule Ingolstadt.

ABSTRACT Environment perception using camera, radar, and/or lidar sensors has significantly improved in the last few years because of deep learning-based methods. However, a large group of these methods fall into the category of supervised learning, which requires a considerable amount of annotated data. Due to uncertainties in multi-sensor data, automating the data labeling process is extremely challenging; hence, it is performed manually to a large extent. Even though full automation of such a process is difficult, semi-automation can be a significant step to ease this process. However, the available work in this regard is still very limited; hence, in this paper, a novel semi-automatic annotation methodology is developed for labeling RGB camera images and 3D automotive radar point cloud data using a smart infrastructure-based sensor setup. This paper also describes a new method for 3D radar background subtraction to remove clutter and a new object category, GROUP, for radar-based object detection for closely located vulnerable road users. To validate the work, a dataset named INFRA-3DRC is created using this methodology, where 75% of the labels are automatically generated. In addition, a radar cluster classifier and an image classifier are developed, trained, and tested on this dataset, achieving accuracy of 98.26% and 94.86%, respectively. The dataset and Python scripts are available at https://fraunhoferivi.github.io/INFRA-3DRC-Dataset/.

INDEX TERMS 3D radar, camera, deep learning, INFRA-3DRC dataset, intelligent roadside infrastructure, labeling, lidar, semi-automatic annotations, smart infrastructure.

I. INTRODUCTION

In recent years, the quality and robustness of environmental perception in terms of road user detection, recognition, and motion prediction using cameras, radar, and lidar sensors have significantly improved. The major contributor to this rapid improvement is the extensive use of deep learning methods [1], [2], [3], which are a subset of artificial intelligence. However, a considerable part of such algorithms fall under the category of supervised learning [4], where the training of algorithms (also known as deep learning models) requires a large amount of annotated data from one or more sensors.

The associate editor coordinating the review of this manuscript and approving it for publication was Salvatore Surdo.

Because of the high uncertainty and unknown patterns in the sensor data, the annotation process becomes challenging. In addition, when two or more sensors' data are annotated, associating information from them leads to additional challenges. As a result, still a large part of the annotations are generated manually which needs huge effort, cost, and human resources. With the increased use of supervised deep learning algorithms, the need to automate the process of sensor data annotation has become crucial. Although fully automating this process of multi-sensor annotations is difficult, partial automation (or semi-automation) is still a way to move forward in this regard to reduce cost and human efforts. There is some work available in the literature dealing with automating the data labeling process [5], [6], [7], [8], [9], [10] but they have limitations,

© 2024 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

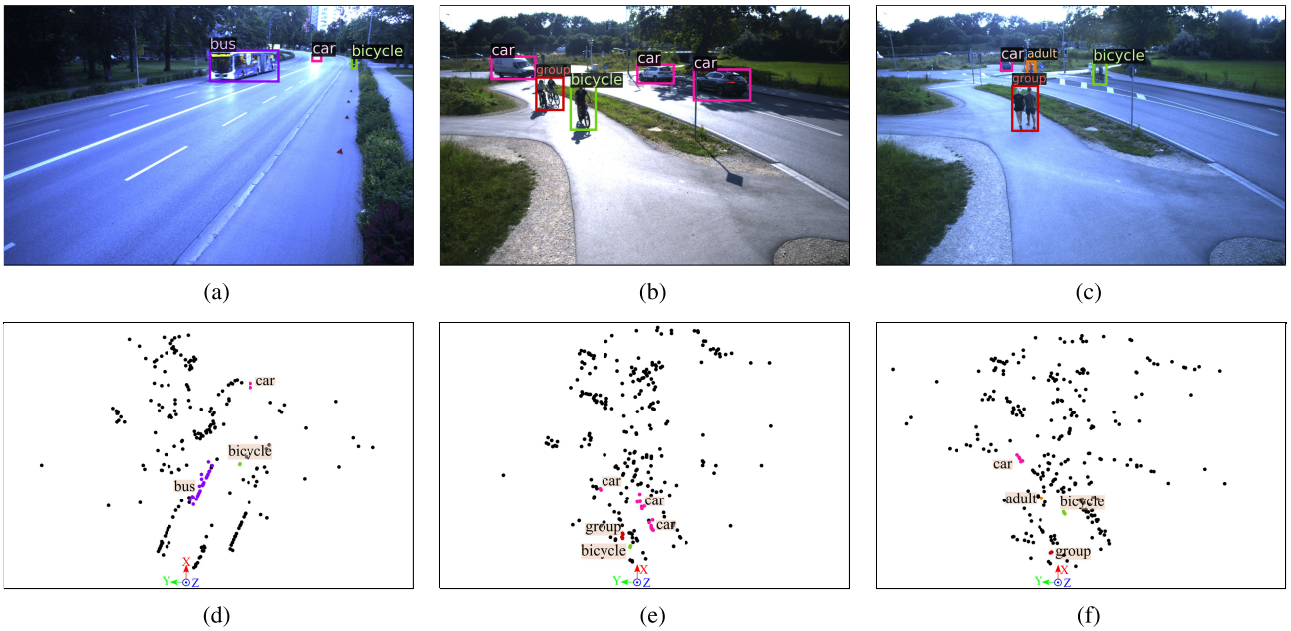

FIGURE 1. Results of proposed semi-automatic annotation methodology. Annotated RGB images are shown on top with a bounding box and object category, and the corresponding annotated 3D radar point cloud (in bird-eye-view) is shown at the bottom. Each point in the radar point cloud is colored according to the category, and black points belong to the background. Please note that the labels on the radar point cloud are only for visualization.

as described later in section II-B. Hence, the work described in this paper focuses on solving this specific issue by proposing a new semi-automatic annotation methodology that specifically focuses on annotating RGB (red, green, blue) camera images and 3D (3-dimension) radar point cloud data.

In a 3D radar sensor, each detection (or point) is associated with range, azimuth angle, elevation angle, doppler speed, and radar cross-section, whereas in 2D (2-dimension) radar, the elevation angle is not available [11]. Hence, 3D radar provides an extra dimension in measurement that considerably increases the spatial resolution and thus the overall point cloud density. Due to this distinct advantage, 3D radar sensors have received acceptance and popularity in many advanced driver assistance systems, autonomous vehicle development, and certain newly published public datasets [10], [12], [13], [14], [15]. However, all these public datasets focus only on vehicle-based sensor setups.

To enhance road safety, along with autonomous vehicles, smart infrastructure-based sensor perception also plays a vital role. In such setups, multiple sensors are mounted at a considerable height along the side of the road to perceive the environment in real-time and to send critical information and warnings to passing road users through a communication channel. Such sensor setups have an elevated view angle, which reduces on-road occlusion to a large extent compared with vehicle-based perception [16], [17]. However, in presently available infrastructure-based setups, as described in Table 3 of [18], projects like [19], [20], and [21] (only some are cited for reference) have used 2D radar sensors along with cameras and/or lidar, but 3D radar sensor that can provide enhanced perception has not yet been explored. To close this gap, the proposed work uses a smart infrastructure-based setup, as described in [17] for the proposed semi-automatic annotation methodology.

In addition, two other challenges in object detection using radar sensors are discussed and novel solutions are proposed. The first challenge is the limited spatial resolution of the 3D radar sensor. Even though the resolution of 3D radar sensors has improved compared to its predecessor, it is still far from the camera and lidar sensors. Hence, when two or more vulnerable road users (VRUs) are either moving or standing very close to each other, the separation of VRUs is very challenging using a radar sensor. To solve this ambiguity in object detection, a new object category - GROUP is proposed which considers such close VRUs as one object in sensor space. More details on this new object category are provided in section V of this paper. The second challenge is the inability of the radar sensor to differentiate between background clutter and static but valid road users, especially when deployed at pedestrian crossings and traffic light junctions as part of smart infrastructure-based units. Hence, a 3D radar background subtraction method is proposed in this work that filters out background clutter in a static setup to a large extent. This method is inspired by the roadside 3D lidar-based background subtraction technique described in [22]. These two solutions are also part of the proposed semi-automatic annotation methodology.

Fig. 1 shows the results of the proposed methodology, where each camera frame is annotated with a bounding

box and object category for valid road users, and the corresponding 3D radar point cloud frame (calibrated and time-synchronized) is annotated point-wise. This means that each point in the radar point cloud is assigned an object category that can be either a valid road user or background, and each road user is identified as a separate instance. The 6 object categories or class labels used in this work are adult (pedestrian), group (described in section V), bicycle, motorcycle, car, and bus. Two more object categories named child and truck to be added in future work.

Additionally, to facilitate research on perception algorithm development with 3D radar sensors in smart infrastructure-based sensor setups, a new dataset, named the INFRA-3DRC dataset, is generated and published using the proposed semi-automatic annotation methodology. This dataset contains annotations of calibrated and time-synchronized 3D radar and RGB mono camera data frames. It also consists of calibrated and synchronized 3D lidar sensor frames. However, the main focus of this work is to annotate 3D radar point cloud data together with camera images; hence, the lidar frames provided in the dataset are not annotated.

This paper is structured as follows: Section II provides the literature review of available datasets in the autonomous driving and smart infrastructure domain. It also includes a review of the available annotation approaches and highlights their limitations. Section III describes the smart infrastructure-based measurement setup used to collect data for this study. It also briefly explains the process of multi-sensor calibration and time synchronization, which are indispensable parts of the complete pipeline. Section IV describes the proposed 3D radar background subtraction method, and section V discusses the importance and definition of the newly introduced object category GROUP, along with some examples selected from the collected data. Section VI describes the semi-automatic annotation methodology proposed in this work for labeling the camera RGB images and the 3D radar point cloud data. Section VII provides statistics of the published INFRA-3DRC dataset and describes the experimental results of the developed and trained image and radar cluster classifiers. Finally, a discussion and conclusion are provided.

A. CONTRIBUTIONS

The main contributions are:

- A novel semi-automatic annotation methodology is developed for RGB mono camera and 3D automotive radar data labeling
- The INFRA-3DRC dataset - an infrastructure-based sensors dataset of RGB camera and 3D automotive radar sensor is published for the research community
- A new 3D automotive radar background subtraction algorithm is developed for static sensor setup to remove clutter.
- A new object category - GROUP is defined for radar-based object detection
- Custom image classifier and radar cluster classifier are developed, trained, and tested on the INFRA-3DRC dataset

II. RELATED WORK

A. AVAILABLE DATASETS

The list of public datasets available in the autonomous vehicle domain is long; hence, in Table 1, only those datasets that include either 2D or 3D radar sensor(s) are listed. In addition, datasets generated using smart infrastructure-based setups are also included in Table 1 but regardless of the availability of the radar sensor.

It is evident from Table 1 that none of the static infrastructure-based datasets available previously in the literature include any 2D or 3D radar sensor for perception. Hence, in this work, a new dataset, named the INFRA-3DRC dataset, is also generated and published. This dataset includes annotated data of 3D automotive radar along with annotated camera RGB images using a smart and static infrastructure-based setup. This is also highlighted in the last row of Table 1.

B. STATE-OF-THE-ART DATA ANNOTATION METHODS

This section highlights the state-of-the-art sensor data annotation methods available in the literature.

The work described in [5] annotates the 3D lidar point cloud by applying foreground and background separation, followed by DBSCAN-based (density based spatial clustering of applications with noise-based) clustering and PCA (principal component analysis). The class labels are transferred using the corresponding camera-based object detection. However, this work does not consider radar sensor data labeling. Work of [6] describes a process to estimate 3D bounding boxes on object proposals generated by tracking the sequence of lidar data. However, it is not clear from the paper how the class labels are generated for the estimated bounding boxes in their automatic annotation framework. In addition, the authors have mentioned the use of proprietary software aiNotate on their dataset website to generate annotations for their work, which is not openly available for other research projects.

The work described in [7] uses neural networks to perform semantic segmentation on camera and lidar images. Then, the method assigns each radar detection to two different labels, one based on the camera and another based on lidar. The best label is determined using the uncertainty-based fusion of both labels. In [8], authors replaced the neural networks of [7] with traditional pipelines, including tracking, to generate labels for radar points. However, the use of two extra sensors to annotate radar points is both computationally and cost-wise expensive. The work proposed in [9] requires instructed traffic participants to wear GNSS (global navigation satellite system) sensors to label the radar points for pedestrians, and cyclists. However, this method is not feasible to annotate different road users in real traffic situations because it requires every traffic participant to be mounted with a GNSS sensor.

TABLE 1. Available datasets.

Dataset	Year	Sensors	Annotation Type	Annotated Frames	Classes	Weather and light	Setup Height (m)	Traffic	Labeling Method
Autonomous driving datasets with 2D radar point cloud.									
nuScenes [23]	2020	LCR	3D	40k	23	SRC/DN	–	USH	M
PixSet [24]	2021	LCR	3D	29k	20	SR/DN	–	USP	M
Pointillism [25]	2020	LCR	3D	54k	–	SF/DN	–	U	M
Dense [26]	2020	LC$_s$R	3D	13.5k	14	SRFS$_n$/DN	–	USHT	M
Autonomous driving datasets with 3D radar point cloud.									
RadarScenes [12]	2021	CR	point-wise	832k	11	SR/D	–	USHT	M
Astyx [10]	2019	LCR	3D	0.5k	7	D	–	SH	S
View-of-Delft [13]	2022	LC$_s$R	3D	8.6k	13	–	–	U	M
RADIal [14]	2021	LCR	2D, seg.	8.2k	–	–	–	USH	–
aiMotive [6]	2023	LCR	3D, 2D	26.5k	14	R/DN	–	UH	S
TJ4DRadSet [15]	2022	LCR	3D	7.7k	8	S/DN	–	U	M
Smart infrastructure datasets.									
BAAI-VANJEE [27]	2021	LC	3D, 2D	2.5k	12	SRC/DN	4.5	U	M
IPS300+ [28]	2022	LC	3D, 2D	14k	7	DN	5.5	U	M
DAIR-V2X_I [29]	2022	LC	3D, 2D	10k	10	SRF/DN	–	UH	M
A9-Dataset [30]	2022	LC	3D	1k	9	D	7	H	M
A9-Intersection [31]	2023	LC	3D	4.8k	10	SRC/DN	7	HI	M
LUMPI [5]	2022	LC	3D, 2D	90k	6	SCH	7	U	S
Ko-PER [32]	2014	LC	3D	4.8k	–	–	5	U	M
Rope3D [33]	2022	C	3D, 2D	50k	13	–	–	U	M
INFRA-3DRC [ours]	**2023**	**LCR-3D**	**2D, point-wise**	**2.7k**	**6**	**S/DN**	**3.5**	**UIP**	**S**

Sensors: L, C, R, C$_s$ stand for lidar, camera, radar, and stereo camera; Annotation Type: 3D, 2D stand for 3D bounding box, and 2D bounding box; Weather and light: S, R, C, F, H, S$_n$, D, N stand for sunny, rainy, cloudy, foggy, hazy, snow, day, and night; Traffic: U, S, H, I, P, T stand for urban, suburban, highway, intersection, parking lot, and tunnel. Labeling Method: M and S denote manual and semi-automatic, respectively. In all columns, "–" indicates that there is no information available.

The Astyx dataset [10] is one of the vehicle-based datasets containing 3D automotive radar point cloud data. The authors used an active learning-based semi-automatic annotation approach in combination with uncertainty-based manual fine-tuning to label the 3D radar point cloud data. However, this approach requires that the initial frames be completely manually labeled to train a deep learning model. In [34], the authors use an image-based YOLO (you only look once) object detector to generate bounding box annotations on the input camera image and a DBSCAN clustering algorithm to generate clusters from 2D radar point cloud. The clusters and image bounding boxes are associated using the Hungarian algorithm after projecting the cluster centroids onto the image plane. However, this method is limited to only 2D radar sensors, and no information is provided regarding the handling of static but valid road users.

Some other works [35], [36], [37], [38], [39], [40] have focused on radar raw data available in the form of RA (range azimuth), RD (range doppler), and/or RAD (range azimuth doppler) cube. Because the proposed work focuses on the processed radar point cloud data, these cases are not within the scope of this work and hence are not explained in detail.

The proposed semi-automatic annotation methodology in this work has clear advantages over the available related work because it tackles the challenges of annotating 3D radar point cloud without relying on any deep learning-based training (that requires heavy computation) or a lidar sensor (in many setup with radar sensor, lidar is not available). Furthermore, it handles the cases of static road user annotation in an infrastructure-based setup, which is a challenging task.

III. MEASUREMENT SETUP AND DATA COLLECTION

This section describes the smart infrastructure setup used for data collection and data generation. It also describes the sensor calibration process and the time synchronization of data frames between sensors.

A. MEASUREMENT SETUP

The measurement setup comprised an RGB mono camera, 3D automotive radar, and 360° automotive lidar sensor. The details of each sensor, mechanical mountings, and electrical connections are described in section IV of [17]. For reference, the same setup is also shown on the left side of Fig. 2 where a tripod is extended to a considerable height on one side of the

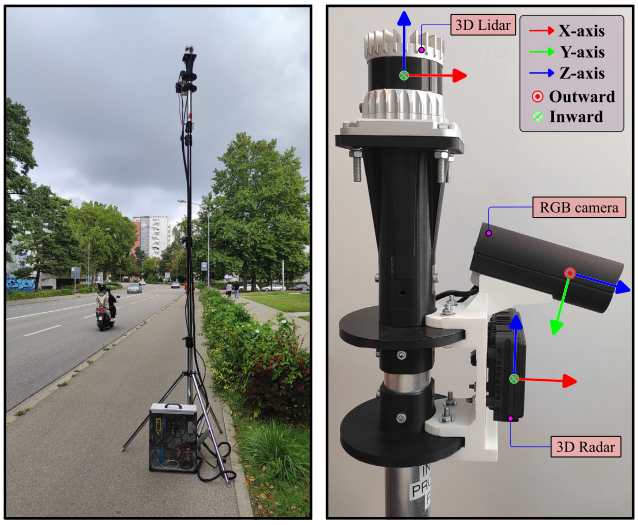

FIGURE 2. Smart infrastructure setup used for measurement (updated from [17].).

road for data collection. The right side of Fig. 2 highlights the coordinate system of each sensor for reference.

B. DATA COLLECTION

Data is collected at different locations including straight roads and crossing junctions with curve roads, in daylight, twilight, and night. During each measurement campaign, the measurement setup is mounted firmly at the side of the road, and the sensors are aligned to adjust the view angle using a height-adjustable tripod, as shown in Fig. 2.

After fixing the setup, on-field calibration is performed using the method described in [41]. This calibration method calculates extrinsic calibration for radar-to-camera, lidar-to-camera, and radar-to-lidar. The ground coordinates for the setup are defined similarly to the lidar coordinates with the origin shifted to the ground (road). The intrinsic calibration of the camera is performed using the checkerboard pattern method of [42] in the laboratory before the measurement campaign. The complete sensor setup is developed using a robot operating system (ROS). During the measurement campaign, data is collected manually in the form of rosbags for a duration of $10 - 15$ seconds each. A graphical user interface (GUI) tool is developed and used to ease the manual collection and sensor data monitoring process.

In the setup, the camera has a frame rate of 30 Hz, and the radar and lidar have a frame rate of 20 Hz each. During post-processing, data from each rosbag is extracted. Camera images are saved as portable network graphics (PNG) files, and radar and lidar point clouds are saved as point cloud data (PCD) files along with their Unix-based timestamps. Using these timestamps, the data frames of radar and camera are synchronized with each other within a delta time of a maximum of 10 milliseconds, and then lidar frames using synchronized camera frames are selected within a delta time of a maximum of 40 milliseconds. All remaining non-synchronized data frames are then discarded. With this approach, approximately 10 Hz of synchronized frames from all three sensors are achieved. All rosbags are post-processed in the same manner and then used for data labeling.

IV. 3D RADAR BACKGROUND SUBTRACTION

In a smart infrastructure setup, radar sensors are mounted at a static position and oriented in a fixed direction. In such conditions, the static environmental view of the sensor remains constant over time; hence, many radar points have the same spatial position (within the given variance due to sensor inherent noise). The majority of these points are generated from static surroundings such as trees, roads, buildings, traffic lights, metal poles, etc. which are not necessary for road user detection. All such points are jointly referred to here as background. Background subtraction of static-mounted radar sensor aims to remove maximum possible background points using appropriate algorithm so that detections from valid road users can be processed optimally. In addition, background subtraction helps to efficiently detect static but valid traffic users, which would have been very difficult without removing background points.

To the best of the author's knowledge, no work has been found regarding 3D radar point cloud-based background subtraction in the literature, and only [22] has described it, but for roadside 3D lidar sensor. The work proposed here is partially inspired by [22] to develop a suitable algorithm for 3D radar-based background subtraction.

3D radar sensor provides data in polar coordinates defined by range, azimuth angle, and elevation angle. Furthermore, each detection is associated with doppler speed and radar cross-section. Background points have near zero doppler speed (but not exactly zero due to noise in radar sensor measurement), and hence, only these points are used for background subtraction. Points with doppler speed ($abs(v) > 0.1\ m/sec$) are filtered out. In Fig. 3, radar point cloud (in bird-eye-view) from the one-time frame is shown, where green points are dynamic points ($abs(v) > 0.1\ m/sec$), and white points are static points. Dynamic points are shown only for visualization and are not included in the background subtraction, as previously stated. Furthermore, the static point cloud shown in Fig. 3 includes data from the background and static (but valid) road users.

The complete process of background subtraction is divided into two parts: background detection and background removal. During background detection, a 3D weighted occupancy polar grid is generated that contains weighted information on the occupancy of background points in the sensor field of view. This step is performed only once for a given fixed view of the sensor setup. If the sensor view changes, this step must be repeated to generate an updated polar grid.

In the second step, the generated 3D weighted occupancy polar grid is used to perform background removal on each radar frame for complete data collected with the same sensor alignment at the same location. Fig. 4, describes a complete

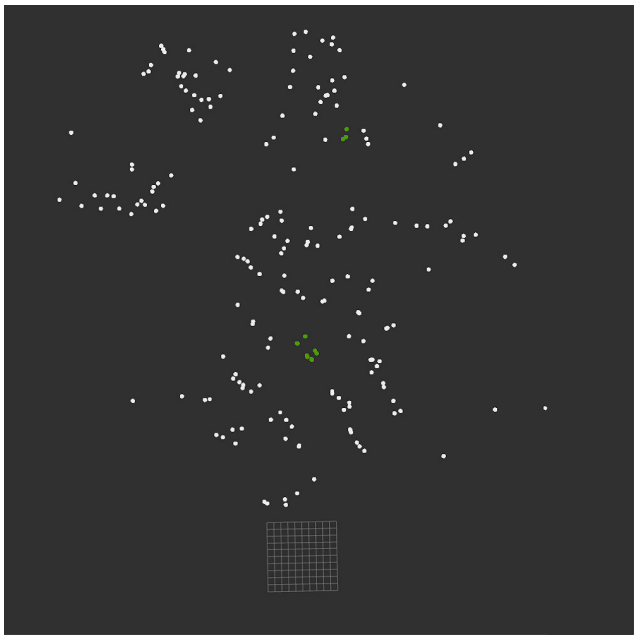

FIGURE 3. Bird-eye-view of 3D automotive radar point cloud before background subtraction. Static points are shown in white and dynamic points are shown in green. Cartesian coordinates are used for visualization.

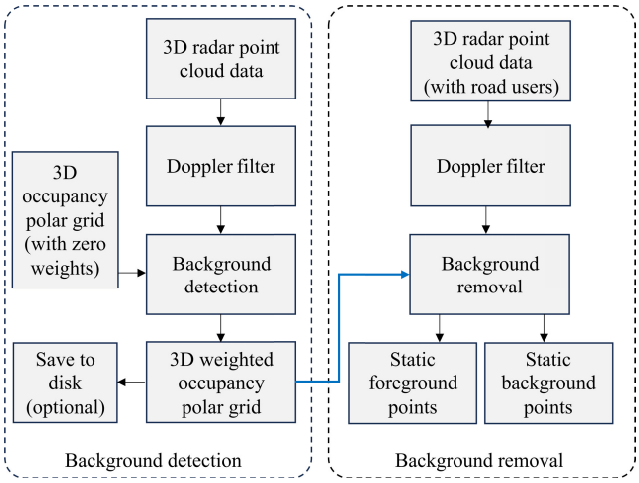

FIGURE 4. Overview of 3D radar background subtraction. Background detection (on the left side) and background removal (on the right side).

process in the form of a block diagram, and the details of each are given later in this section.

A. BACKGROUND DETECTION

For a given location, after setting up the sensors, a small scene is recorded only with radar data for a few seconds. The duration of these data depends on the frame rate of the sensor and the required minimum frames for optimum background detection. This depends on multiple factors such as sensor type, location, environment, etc. For this work, data with a minimum of 600 consecutive radar frames is selected after some experiments. If any valid static road user is present during this data collection, those locations might get detected wrongly as background. Hence, the ideal choice during data collection for background detection is that no valid road user should be available in the environment; however, if this is difficult, then collecting data with only dynamic road users is recommended because dynamic points are filtered out before performing background subtraction.

For background detection, the complete field of view (or required field of view) of the sensor is divided into a 3D polar mesh grid. The dimension of a grid cell is taken as per the resolution of the sensor in each dimension, i.e., range resolution R_{res}, azimuth angle resolution A_{res}, and elevation angle resolution E_{res}. For the radar sensor used in this work, this information is available in [43]. Furthermore, the total grid cells in each dimension are calculated as per equation (1), where R_{tc}, A_{tc}, and E_{tc} are the total grid cells in range, azimuth, and elevation.

$$\begin{aligned} R_{tc} &= \frac{R_{max} - R_{min}}{R_{res}} + 1 \\ A_{tc} &= \frac{A_{max} - A_{min}}{A_{res}} + 1 \\ E_{tc} &= \frac{E_{max} - E_{min}}{E_{res}} + 1 \end{aligned} \tag{1}$$

In equation (1), R_{max}, A_{max}, and E_{max} are the maximum values, and R_{min}, A_{min}, and E_{min} are the minimum values of range, azimuth, and elevation angles, respectively that can be measured by the radar sensor. The total grid cells formed in the sensor field of view are calculated using equation (2), where FoV_{tc} is the total grid cells in the sensor FoV. These total grid cells also indicate the total possible locations that can be associated with either background or foreground.

$$\mathrm{FoV_{tc}} = R_{tc} A_{tc} E_{tc} \tag{2}$$

Once the 3D polar grid is created, each grid cell is assigned zero weight, which means that there is no background occupancy. The first radar frame is taken, and all detection values are assigned to grid cells according to their range, azimuth, and elevation angle. Then, for all grid cells where detection is associated, the weights are incremented by 1, and the weights of all non-associated cells remain unchanged. The same process is performed on all radar frames consecutively, and for every frame, the weights of the associated grid cells are increased by 1 with respect to the previous value. For example, if for a particular grid cell, a point is associated in 200 frames, that grid cell will weigh 200. The final 3D polar grid with each cell associated with a certain weight is referred to as a 3D weighted occupancy polar grid, which is the result of the background detection algorithm. It can be optionally saved to disk for later use

B. BACKGROUND REMOVAL

The pre-calculated 3D weighted occupancy polar grid is loaded from the disk or the file, and a weight threshold value (one hyper-parameter) is selected that decides whether a particular cell in the 3D polar grid is considered as

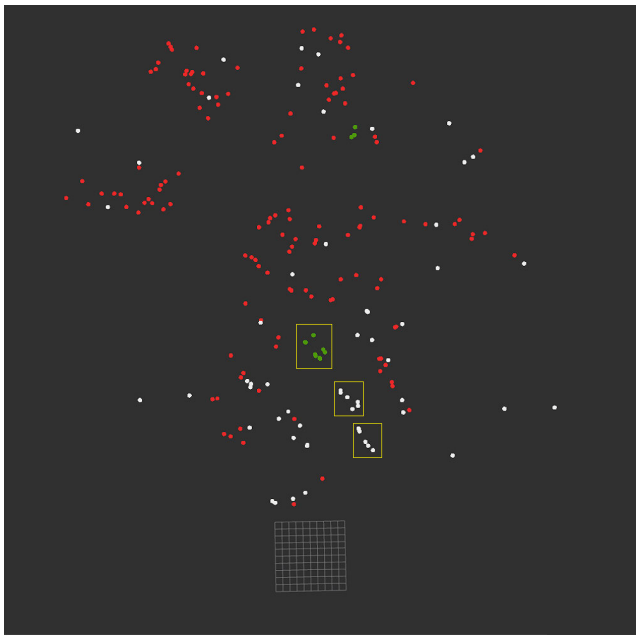

FIGURE 5. Bird-eye-view of 3D automotive radar point cloud after background subtraction. From the static points, red points are associated with background points, and white points are associated with foreground points. Dynamic points are shown in green. Cartesian coordinates are used for visualization.

background or foreground. In this work, a value of 10 (no unit) is selected after the experiments. This means that for all 3D polar grid cells if the weight of the cell is greater than 10, it is considered the background point, and if it is less than 10, it is considered the foreground point. Using this value, the 3D weighted occupancy polar grid is converted to a 3D binary occupancy polar grid.

Background removal is applied to each radar frame used for object detection. For this purpose, the radar frame is first filtered to separate dynamic and static points. Then, all the static points are assigned to the 3D binary occupancy polar grid. If the point is associated with the cell with the value true (or 1), it is considered a background point, and if the point is associated with the cell with the value false (or 0), it is considered a foreground point. Hence, the background removal algorithm outputs static points separated as either background or foreground points. The same process is applied to other radar frames.

Fig. 5, shows the results of background subtraction applied to the radar frame highlighted in Fig. 3. The dynamic points (shown in green) remain unchanged, whereas all the static points are categorized either as background points (shown in red) or as foreground points (shown in white).

Fig. 6 shows the camera image calibrated and time synchronized to the radar frame given in Fig. 3, and Fig. 5. From Fig. 6, it is evident that only three cars are available in the sensor field of view, which are marked with yellow boxes. The same is also marked with yellow boxes in the radar point cloud in Fig. 5. From the radar point cloud, it is confirmed that only one car is moving (having points in green), while

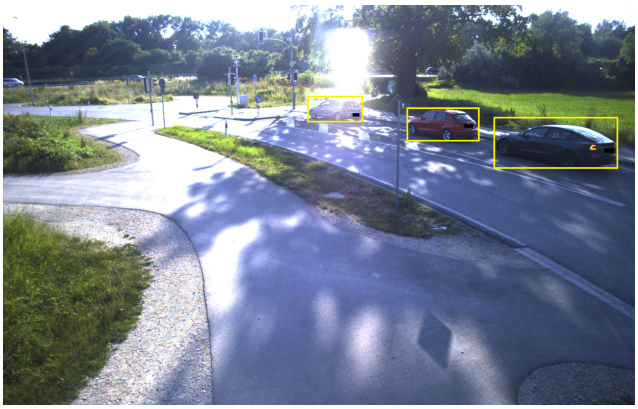

FIGURE 6. RGB camera image for reference with radar point cloud of Fig. 3, and Fig. 5.

the other two cars are standing (having points in white) near the traffic light junction. After background subtraction, valid traffic users are successfully assigned as foreground points, and maximum clutter is assigned to the background.

As radar sensor data is noisy in nature, one cannot remove all the clutter using background subtraction. However, as shown in Fig. 5, the large number of background points are successfully removed by this algorithm.

In this background subtraction, RCS (radar cross-section) is not used to differentiate between static background and static traffic users, because from the collected data, the distribution of RCS does not show a clear difference between these point clouds. However, further analysis will be conducted using the collected data in the future. Moreover, this algorithm will be studied and adapted for different weather conditions, such as rain, snow, and fog, and for corner cases, such as when a person is sitting or lying down on the road, as part of the future work.

V. NEW OBJECT CATEGORY-GROUP

The low spatial resolution of the 2D automotive radar sensor is improved considerably in the 3D version [10], but still, it is not comparable with camera and lidar sensors. Hence, in many situations, specifically with VRUs (vulnerable road users) that include bicycles, adults, and children, it is very difficult for radar sensors to differentiate each road user separately when they are moving or standing very close to each other. This leads to ambiguity in VRU detection with radar sensors. The proposed new category - GROUP in this work aims to solve this problem wisely for object detection. Please note that this object category is not for other road users such as cars or buses. These are detected as separate entities by the sensor. Moreover, the GROUP category is defined here because the 3D radar sensor used in this work does not provide micro-doppler measurements. With the help of micro-doppler data, closely moving VRUs can be differentiated to a certain extent, but closely standing VRUs are still very challenging to detect separately.

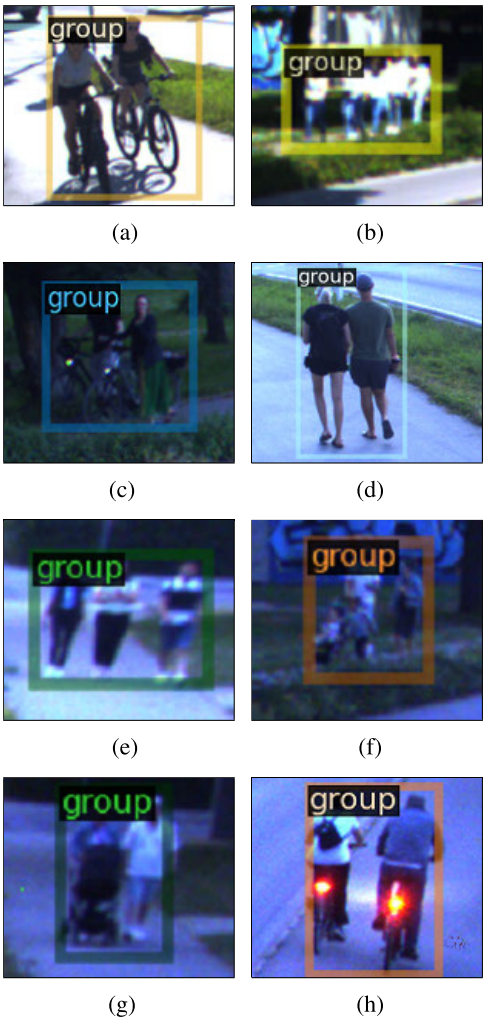

FIGURE 7. Different sample images highlighting the various conditions of new category - group.

The main aim of any real-world object detector (in 2D or 3D) is to either know where the object is located or to feed it to the tracking algorithm to obtain the trajectory of the road user. When two or more VRUs are moving or standing very close to each other, they can be considered as one virtual physical entity from the sensor's perspective, even though they are different physical bodies in the real world. In infrastructure-based and vehicle-based perception, the main aim of object detection and motion prediction is to determine the drivable area for the vehicle during path planning.

As an example, when two people are walking side by side with negligible gap between them (normal and frequent situation on the road), detecting them as two different road users provides no performance improvement compared to when both are considered as one object. As there is no drivable area between them, from detection as well as from a tracking point of view, it is efficient to consider them as one object.

In the literature, [12], a category called the pedestrian group is defined, but it is limited to only pedestrians. Situations with multiple bicycles and pedestrians with bicycles are not considered and not defined. Hence, the category - "GROUP" is proposed in this work that also considers situations including pedestrians and bicycles. It is defined as when two or more persons (adult/child) and/or two or more bicycles are either moving or standing close to each other such that there is not sufficient drivable space for any other road user to pass through them.

Fig. 7 highlights some examples taken from real data collected using an infrastructure-based setup. Please note that here examples with images are shown because for the annotation work, camera images are initially manually labeled for category group and then the radar point cloud is annotated point-wise using the semi-automatic annotation methodology described later in this work. There are many other combinations on the road that fall under this category, but it is not realistic to show all of them here.

VI. METHODOLOGY

The proposed semi-automatic annotation methodology labels the camera images with 2D bounding boxes and object categories (also known as class labels) and performs instance point-wise segmentation of the 3D radar point cloud. This means that each detection or point of the radar point cloud is classified as one of the required object categories or background. Furthermore, each instance of the same object is separately identified. A high-level block diagram of this methodology is shown in Fig. 8.

The input comprised an RGB camera image and a 3D automotive radar point cloud of a calibrated and synchronized time frame. The camera image is fed into the image pre-processing module that generates detections in the form of bounding boxes (including class and score) and object masks. Similarly, the radar point cloud is fed into the radar pre-processing module that generates clusters for both dynamic and static road users. The detections from both sensor frames are provided to the auto-labeling module. Once the frames are processed by the labeling algorithm, each frame is manually validated, and the required frames are then classified into correct frames, frames for label change, and frames to manually label. Frames selected for label change are those frames in which the image processing module detected the object correctly but classified it incorrectly. The frames selected for manual labeling are frames with corner cases and objects with special classes, as described later in this section.

Frames selected as correct frames are directly used to create annotations in JSON (javascript object notation) file format for each sensor. The camera annotation JSON file contains the object bounding box locations and the object class or category. The radar annotation JSON file contains point-wise class and instance information for a complete radar cloud. Please note that the object mask generated during image pre-processing is used internally for the labeling process, but it is not part of the final annotations. The

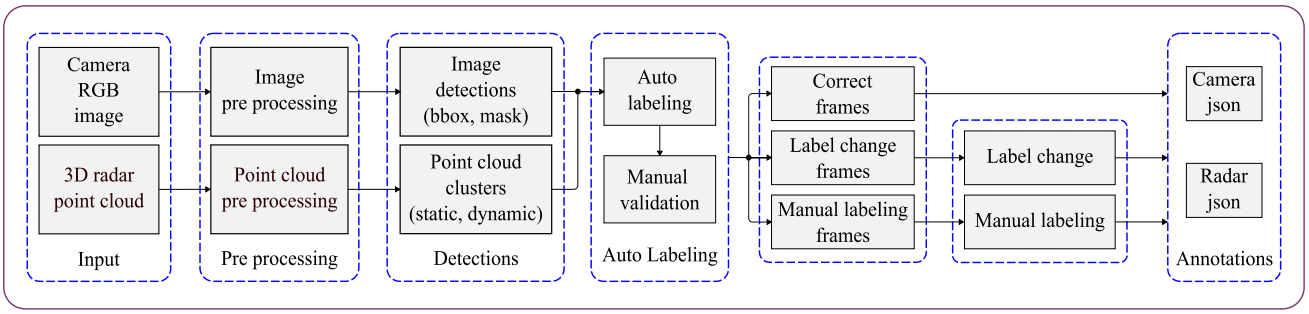

FIGURE 8. Proposed methodology for semi-automatic annotation of camera images and 3D automotive radar point cloud data.

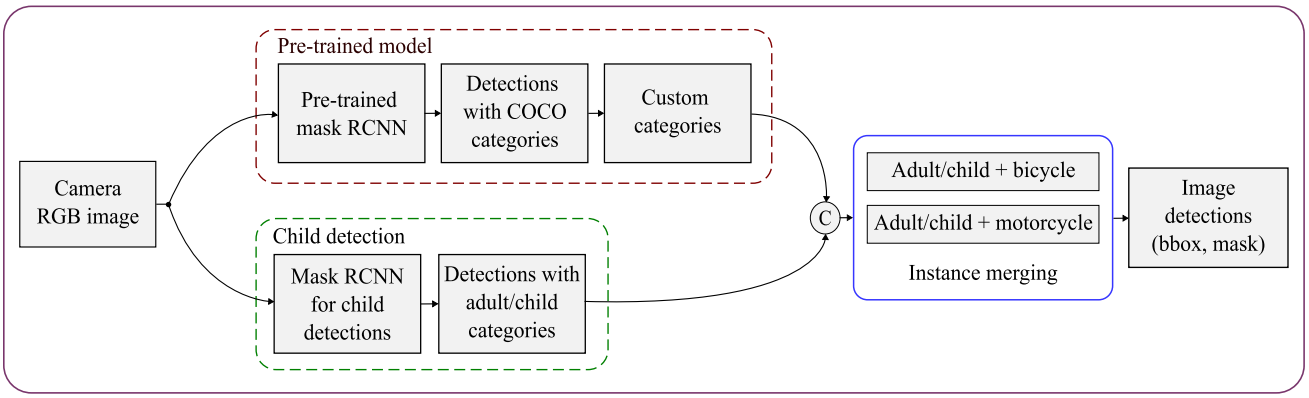

FIGURE 9. Image Pre-processing Pipeline.

frames selected for the label change and manual labeling are processed separately before generating annotations.

In this methodology, most frames are auto-labeled, and only corner cases need to be handled manually. Hence, it is called a semi-automatic annotation methodology.

A. INPUT

Input to the semi-automatic annotation pipeline is a camera RGB image with a resolution of 1920 x 1216 pixels and a 3D Radar point cloud comprised of multiple radar detections (also known as points). Each radar point is associated with range (in meters), azimuth angle or horizontal angle (in radians), elevation angle or vertical angle (in radians), doppler velocity (in meters/second), and RCS (in decibel/square meter).

B. IMAGE PRE-PROCESSING PIPELINE

The image pre-processing pipeline is shown in Fig. 9. In this pipeline, the camera image is fed into the pre-trained mask R-CNN (region-based convolutional neural network) [44] which generates bounding boxes, masks, object categories, and the confidence of detection for each object defined as per the pre-defined categories of the COCO (common objects in context) dataset. The COCO-based categories are mapped to custom categories to remove unwanted categories and add required object categories.

In many instances, when a person is riding a bicycle, the pre-trained mask R-CNN generates two bounding boxes, one for the person and one for the bicycle. Similarly, when a person is driving a motorbike or motorcycle, two separate bounding boxes are generated for the person and motorcycle. Hence, IOU (intersection over union) based instance merging is used to combine such cases into one bounding box for both a person and a bicycle or motorcycle. Some sample images of instance merging from the collected data are shown in Fig. 10.

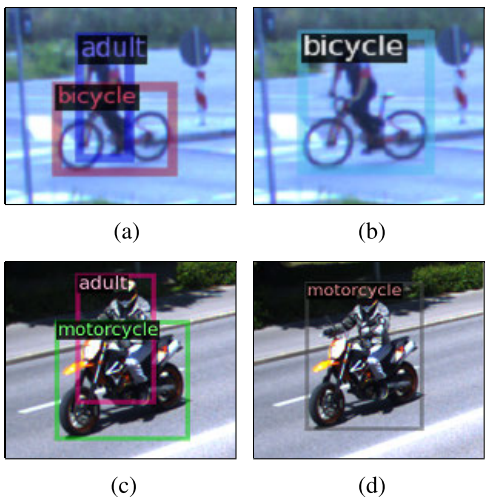

FIGURE 10. Examples of instance merging in camera images. The images on the left are the original annotations from the pre-trained network, and the corresponding images on the right are after instance merging.

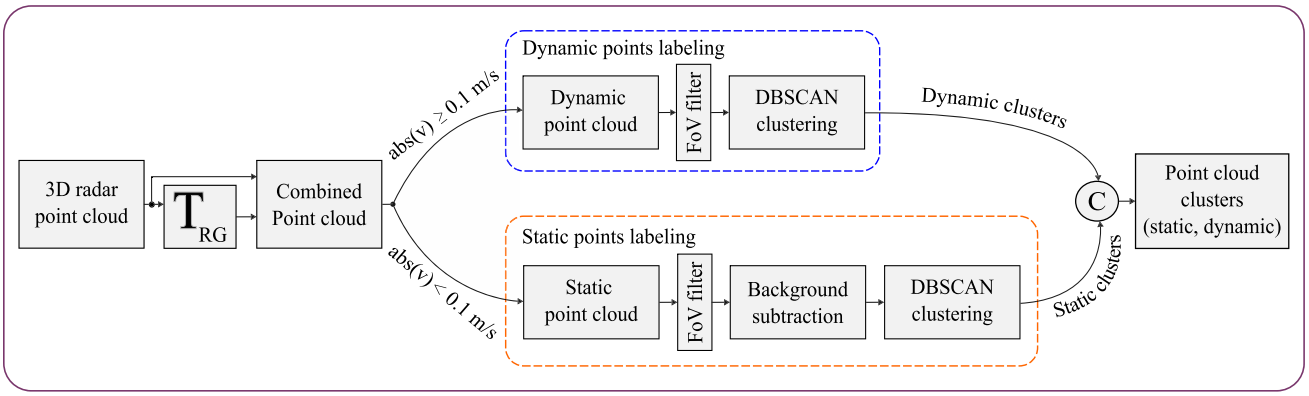

FIGURE 11. Radar pre-processing pipeline.

A separate child detector model, as described in [45] is pre-trained using transfer learning with mask R-CNN to detect adults and children separately. This is also used in parallel to the original mask R-CNN model, and the instances of the person from the original model are replaced by adult or child categories generated by this model. Then the final list of detections associated with 2D bounding boxes, masks, object categories, and detection scores is fed into the auto-labeling process.

C. RADAR PRE-PROCESSING PIPELINE

The radar pre-processing pipeline is shown in Fig. 11. The 3D radar point cloud data of one frame comprised multiple detections measured in polar coordinates. It is then converted into cartesian coordinates and then transformed into the ground plane of the smart infrastructure setup using the transformation matrix T_{RG} as shown in Fig. 11. The values of the transformed and original radar point cloud data are combined for further processing.

By working principle, it is difficult to differentiate between a static road user (say a car or person) and background clutter points. This makes the annotation of static but valid traffic users challenging for radar data. Hence, at first, the complete point cloud is separated into dynamic and static point clouds using a doppler speed filter with an absolute value of 0.1 m/sec. Separate processing pipelines are then used to process the detections of each type. To generate dynamic road user clusters, a field-of-view filter is applied to remove unwanted detections from the far field, and then DBSCAN-based clustering is applied with parameters ($eps = 3$, $minimum\ points = 2$). For static points, after performing a similar field-of-view filter, background subtraction is applied to remove the maximum possible clutter. The process of background subtraction of 3D radar point cloud is described in section IV. After background subtraction, only the foreground points are fed into DBSCAN clustering with the same parameters as those used for dynamic clustering to generate static clusters. Both dynamic and static clusters are then added to the auto-labeling process.

D. AUTOMATIC LABELING AND ANNOTATION GENERATION

The algorithm for the automatic labeling of the camera RGB images and 3D radar point cloud is shown in Fig. 12. The complete process is divided into a total of six stages that are executed one after another. Before starting with stage one, a list of radar clusters (dynamic and static) and a list of image detections (objects with mask, bounding box, class, and score) are generated by executing a 3D radar pre-processing pipeline and a camera image pre-processing pipeline on the synchronized sensor frames of the radar and camera, respectively, as shown in the top part of Fig. 12. This generated output is used as input to stage one of the auto-labeling algorithm, which provides separate lists of non-associated and associated radar clusters and image detections. The associated data is stored and used for annotation generation, while non-associated data is given further to stage two. This is repeated in the next stages.

In Fig. 12, only stage one is described because stages two to six follow almost the same logic as stage one, and only blocks highlighted with circular numbers as 1, 2, and 3 changes. In stage one, to find the clusters associated with image detections, only dynamic clusters are used, and for association, image masks are used. The rest of the algorithm is self-explanatory in the given flow chart. The type of input used in each of the three blocks in all stages is given in Table 2

TABLE 2. Data input used in different stages of labeling algorithm described in Fig. 12.

stages	block 1	block 2 and 3
1	dynamic cluster id $< N_d$	binary mask
2	dynamic cluster id $< N_d$	bounding box
3	dynamic cluster id $< N_d$	expanded bounding box
4	static cluster id $< N_s$	binary mask
5	static cluster id $< N_s$	bounding box
6	static cluster id $< N_s$	expanded bounding box

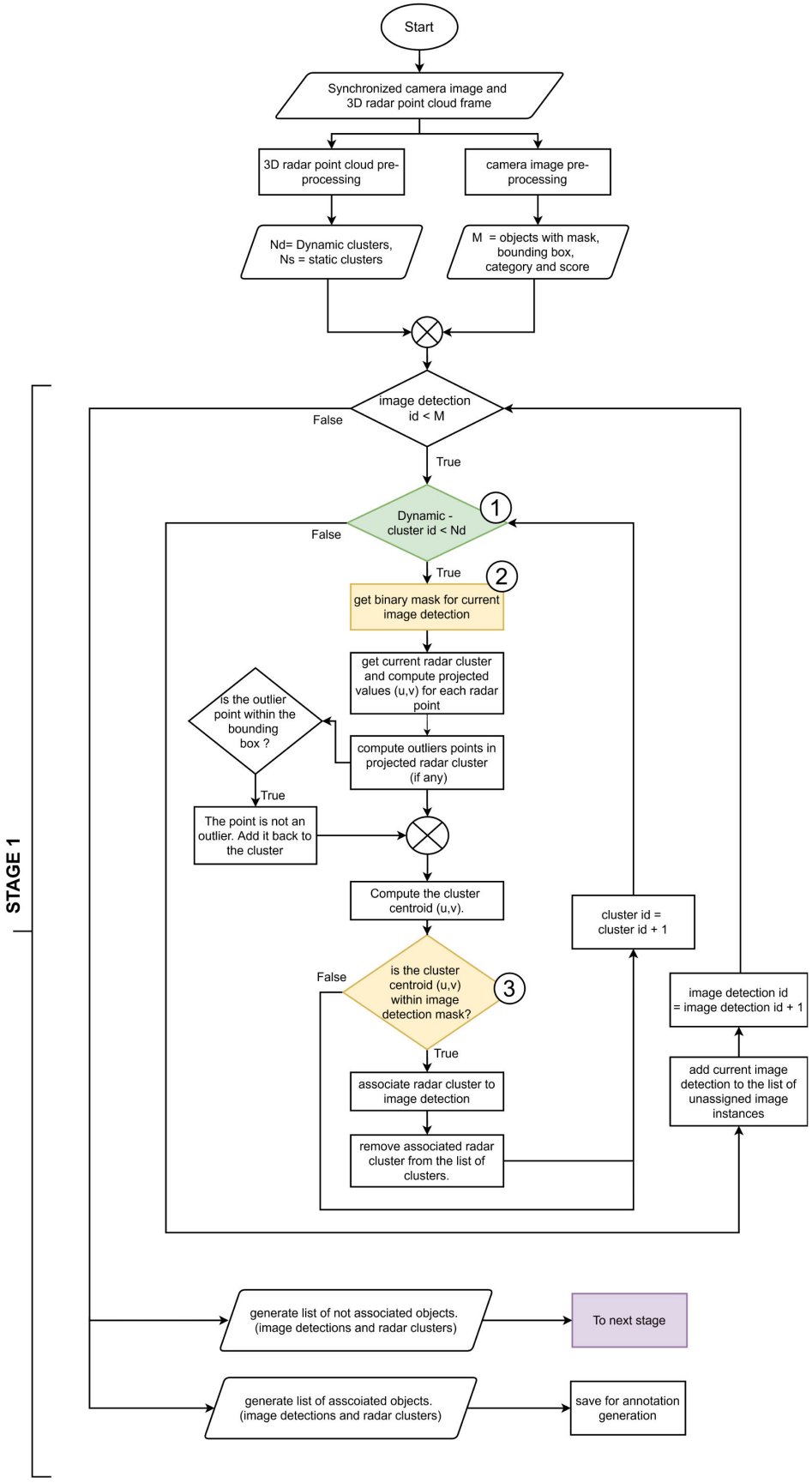

FIGURE 12. Automatic labeling algorithm.

where N_d and N_s are the number of dynamic and static radar clusters, respectively, fed into each stage of the algorithm.

In stages one, two, and three, only dynamic radar clusters are used as input and static clusters are processed in stages four, five, and six. To associate clusters with image detection, three different inputs are used sequentially. At first, image masks are used, then bounding boxes, and then expanded bounding boxes are used. The expanded bounding boxes are generated by uniformly expanding the original bounding boxes by 25%. Due to calibration and time synchronization errors, the radar centroid is sometimes unable to associate with the image mask and bounding box, especially for small-sized image-space objects. Therefore, using an expanded bounding box in such cases increases the association and overall quality of the labeling algorithm.

It is not necessary that for every frame, all six stages are performed. It depends on the type of radar clusters available and the list of non-associated objects left after each stage. If at a certain stage, all objects are associated, further processing is not required. Similarly, if only dynamic radar clusters are available in a frame, only the first three stages are performed.

E. MANUAL VALIDATION OF ANNOTATED FRAMES

Manual validation of annotated frames is a relatively simple process. For every annotated frame, it is checked whether the frame has valid annotations or not. For this purpose, each camera frame is visualized with object bounding boxes, and then radar points are projected on the camera image with their clusters and instances. During the process of validation, each frame (camera + radar) is classified as a correct frame, a label change frame, or a manual label frame.

Correct frame implies that the automatic labeling framework output is satisfactory for that frame and can be used directly. Label change frame occurs when one or more object categories of the correctly detected bounding box from camera pre-processing module are wrong. In this case, it is comparatively simple to change the category in JSON files in a post-processing step. A manual label frame is a frame where automatic labeling has failed for one or more objects. This can be due to one or more reasons, such as when the camera image is not visible enough to detect an object(s), any new object category is defined that is not part of pre-trained mask R-CNN output, like a new category - GROUP, defined in this work or when only one radar point is reflected from the object that is not clustered and hence didn't associate during auto-labeling algorithm. Such frames are then manually labeled in camera images and passed again through the auto-labeling pipeline to generate radar annotations.

F. PROCESSING OF LABEL CHANGE FRAMES

In this process, the frames are manually checked to determine the correct object category (or label) and then updated during the creation of annotation JSON files for camera and radar data. For example, in some images, a bicycle is incorrectly labeled as an adult (pedestrian) by a state-of-the-art object detector. In this step, such labels are changed to correct labels such as a bicycle. Similarly, in some images, a van is classified as a truck, but for this work, it is then changed to a car.

G. PROCESSING OF MANUAL LABELING FRAMES

In manual labeling, bounding boxes are hand-crafted from camera images by humans. To simplify the process, radar clusters from dynamic and static road users generated from the radar pre-processing pipeline are projected onto the corresponding camera image using the projection matrix T_{RC} as shown in Fig. 13. This image with projected radar clusters acts as a reference image for manual labeling. Then, the actual camera image is loaded into the open source python-based labeling tool, Labelimg [46]. Bounding boxes are manually created on the valid road users in the image, and a reference image is used to identify the valid road users. Once the bounding boxes are created, the annotated camera image and corresponding radar clusters are fed back into the auto-labeling module, which generates the annotations for the camera and radar data.

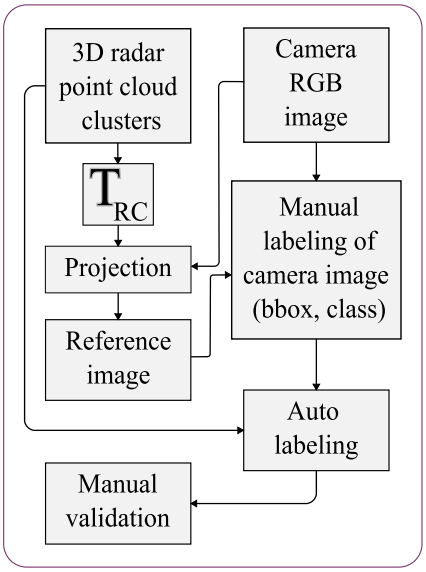

FIGURE 13. Processing pipeline of manual labeling in corner cases.

VII. EXPERIMENTS AND DISCUSSIONS

To validate the described semi-automatic annotation methodology, a large amount of data (RGB mono camera images and 3D radar point cloud frames) is annotated for sensor data fusion development, which will be separately published later. However, a considerable subset of these annotated data is published in the public domain as part of this work and is referred to as the INFRA-3DRC dataset. Details of this dataset including all relevant statistics are provided later in this section.

In addition, to prove that the given dataset generated using the proposed methodology is suitable for the research and development of different perception algorithms using

deep learning methods, a deep learning-based radar cluster classification model and an image classification model are developed, trained, and tested using this dataset.

A. INFRA-3DRC DATASET

The dataset comprised 25 scenes recorded using the smart infrastructure setup described in section III. These scenes are recorded at three different locations. The first location is a pedestrian crossing junction with traffic lights and a curved road, the second location is a multi-lane bidirectional straight road, and the third location is an open parking space. Apart from different locations, data is collected during daylight, twilight (in the evening time), and night.

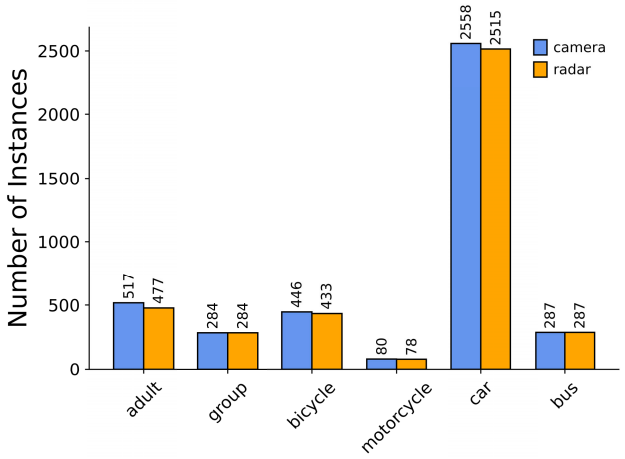

FIGURE 14. Instance-wise distribution of object categories in camera and radar.

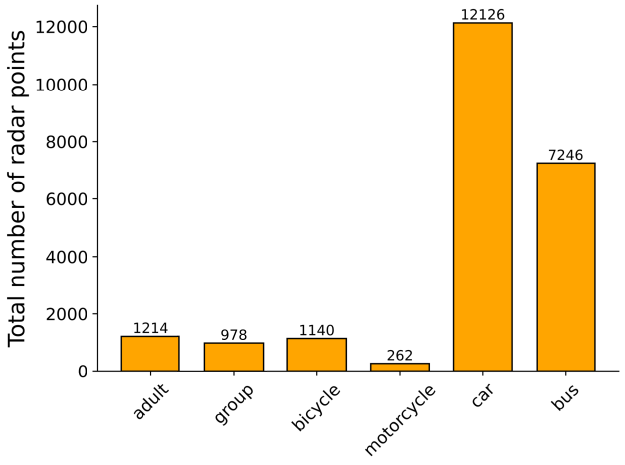

FIGURE 15. Distribution of total radar points labeled in each object category.

The dataset contains a total of 2, 768 annotated frames, each of the RGB camera and the 3D radar, and the same number of non-annotated lidar frames. Instances of six object categories are provided: adult, group, bicycle, motorcycle,

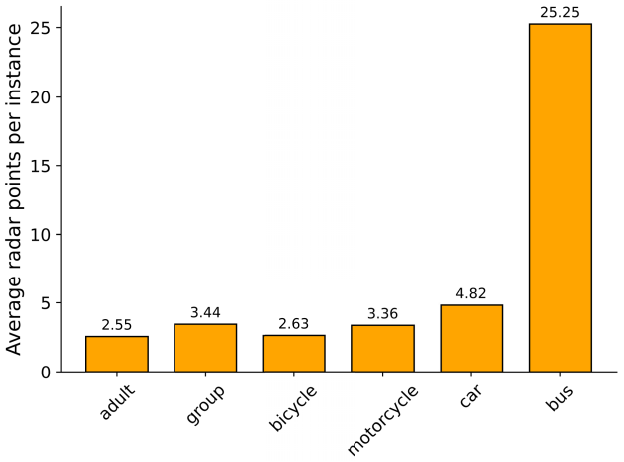

FIGURE 16. Distribution of the average number of radar points available in one instance of each object category.

car, and bus. To generate annotations, camera, and radar frames are input to the semi-automatic annotation framework described in this work, and annotations are stored in JSON files. Furthermore, to comply with the European general data protection regulation (GDPR) [47] of data privacy, clearly visible human faces and vehicle number plates in camera images are anonymized using state-of-the-art algorithms known to the best of the author's knowledge. In each scene, a unique track ID is also associated with every valid traffic user using a separate visual 2D multi-object tracking algorithm that is not part of the described work. This is added in order to facilitate the multi-object tracking algorithm development along with object classification, detection, and segmentation algorithms using this dataset.

Fig. 14 shows the total number of instances of each object category of the camera and radar in the complete dataset. In some object categories, the number of instances in radar is less than that in camera instances because, in a few instances, the radar sensor has no points reflected from the object. This is inherent to radar sensors because in certain cases, either due to high noise in the reflected signal or due to inappropriate angle formation between the object and the sensor, some radar reflections do not qualify as valid detections. Therefore, in such cases, only camera instances are included in the annotation file. The camera data has a total of 4172 instances, and the radar data has a total of 4074 instances of all object categories. Hence, in the complete dataset, 98 instances (2.34% instances) have only camera annotations.

Fig. 15 shows the distribution of the total number of radar points labeled in each object category. In the complete dataset, 22, 966 radar points are labeled for valid traffic users. Fig. 16 shows the distribution of the average number of radar points available in one instance of each object category. The actual number of points in different instances of the same object category can have a large deviation from the average points. For example, in the category of cars, even though

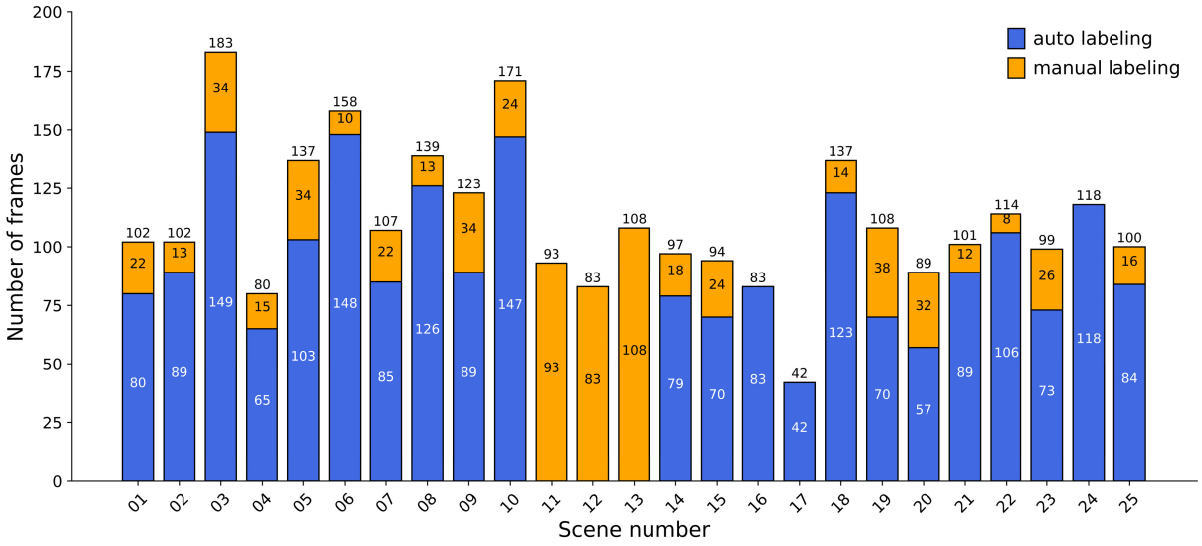

FIGURE 17. Distribution of the auto-labeled and manually labeled frames in each scene of the dataset.

the average radar point is 4.82 (approximately 5 points per instance) but when the car is seen by the radar sensor from the side, the number of points is 10 or more.

Fig. 17 highlights the results of the proposed semi-automatic labeling methodology for the published dataset. For each scene, the total number of auto-labeled and manually labeled frames is given. From a total of 2, 768 frames of all 25 scenes, 2075 frames are auto-labeled and 693 frames are manually labeled. It means 75% of the total frames are auto-labeled using this methodology. Further, if the scenes 11, 12, and 13 of Fig. 17 are removed from the calculation because these three scenes contain the newly proposed object category GROUP, which requires complete manual labeling, then the contribution of auto-labeling in the published dataset reaches 85%.

B. IMAGE CLASSIFIER

A deep neural network-based image classifier is developed and trained for 6 object categories using the annotated dataset described in section VII-A. The architecture of the classifier model is shown in Fig. 18. It has a total of 204k learning parameters.

The distribution of object instances in annotated image data is given in Fig. 14. From this distribution, 10% of the instances from each object category are randomly selected for validation and 15% for the test. Because the number of instances of motorcycles is very low compared with other categories, various augmentation techniques such as horizontal flip, brightness, contrast, and rotation are used to increase the number of instances of motorcycles.

Then, training is performed for 40 epochs using the parameters highlighted in Table 3. The trained model is used on a test set to generate predictions that provide an accuracy

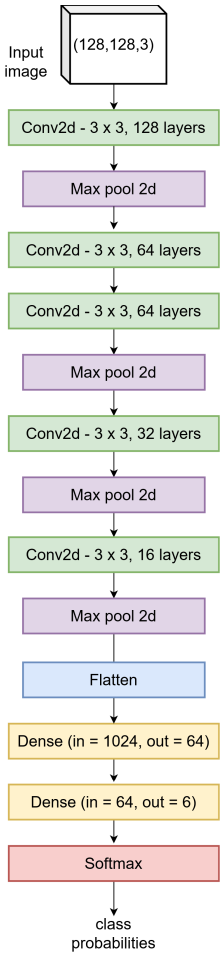

FIGURE 18. Neural network architecture of image classifier.

of 98.26%. Fig. 19 highlights the confusion matrix generated using the test set.

TABLE 3. Image classifier model parameters used during training.

Parameter	Value
batch size	8
loss	cross entropy
optimizer	Adam
learning rate	0.0001
input size	(128, 128, 3)
output classes	6

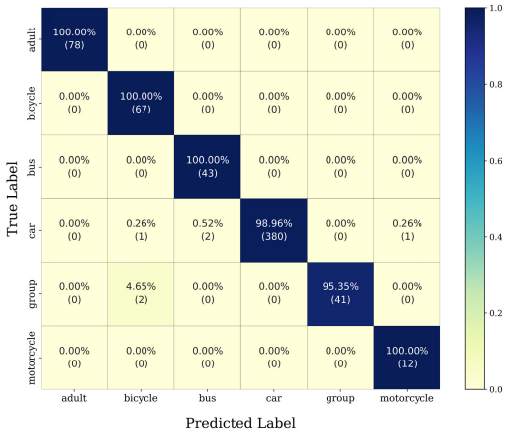

FIGURE 19. Confusion matrix of the trained image classifier.

TABLE 4. 3D Radar clusters classifier model parameters used during training.

Parameter	Value
input features	x, y, z, *range_rate*, *rcs*
batch size	4
loss	cross entropy
optimizer	SGD
initial learning rate	0.1
end learning rate	0.001
learning rate scheduler	linear
output classes	6

C. 3D RADAR CLASSIFIER

For the 3D radar cluster classification task, the entire dataset is split such that 70%, 10%, and 20% of instances from each class are randomly selected for training, validation, and test sets, respectively. To mitigate the risk of inefficient training caused by a class imbalance in the training dataset, a class weighting scheme is used in the cross-entropy loss function. In this way, the loss of samples that belong to the minority class in the training dataset gets a higher weight, enabling the network to focus more on learning the under-represented classes using only a few training samples.

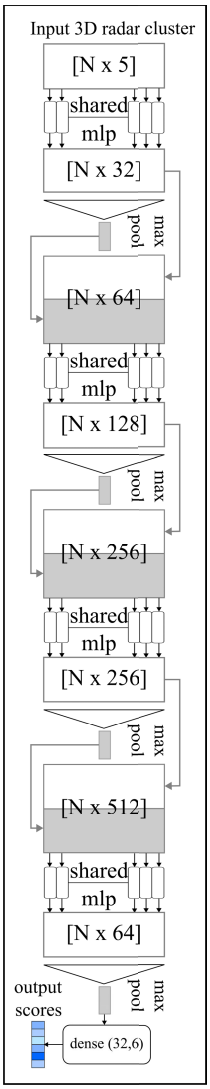

FIGURE 20. Neural network architecture of radar classifier.

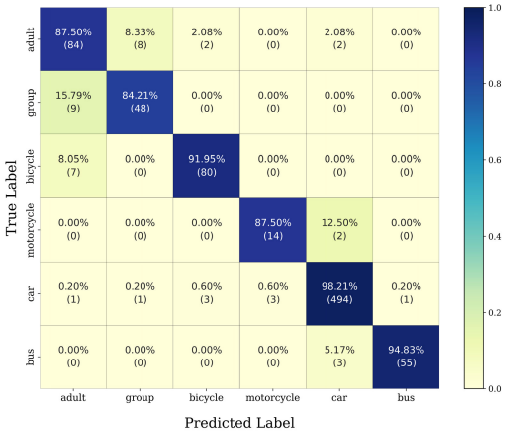

FIGURE 21. Confusion matrix of the trained radar classifier.

Fig. 20 shows the architecture of the developed neural network classifier. It contains 109k trainable parameters for classifying the 3D radar point cloud clusters. The network

is partly inspired by pointnet [48]. Because the number of points N varies across clusters, shared MLPs (multi-layered perceptrons) are used for local feature extraction. Each MLP is followed by a relu activation function, except for the final dense layer, which uses softmax activation to generate class-wise object probabilities. At each stage, the network also captures global features using a max pooling operation and fuses them with local features generated using shared MLP.

The network is trained for 50 epochs using the parameter values highlighted in Table 4. The network takes a cluster of 3D radar points as an input and outputs class probabilities for 6 classes. Each feature of the input cluster is normalized using statistics generated from the training dataset to ensure stable training of the model. After training, the model performance was evaluated on the test set for generating the confusion matrix shown in Fig. 21. The accuracy of the model on the test set is 94.86%.

In Fig. 19 and 21, the total percentage of true positives (correct classification) of the category ''GROUP'' is slightly less than other categories because INFRA-3DRC dataset contains less number of instances in this category and hence model misclassified some instances into adults or bicycles. This can be well improved by feeding more data.

VIII. CONCLUSION

A semi-automatic annotation methodology to annotate RGB mono camera images and 3D automotive radar point cloud frames in a smart infrastructure-based sensor setup is presented in this work. To validate the work, a new dataset, named the INFRA-3DRC dataset is generated within the scope of the work and published using this methodology, where 75% of the total frames were annotated automatically without human intervention. Further, an image classifier and a radar cluster classifier are developed, trained, and tested on this dataset, resulting in an accuracy of 98.25% and 94.86% respectively. This indicates that the described methodology reduces human efforts, cost, and time required for data labeling. Further, it is well suitable to generate custom datasets for camera and radar sensors to develop AI models for classification (presented in this work), object detection, segmentation, multi-object tracking, etc. There are some corner cases where manual labeling work is still required, such as crowded traffic scenes where distant objects overlap in the image plane, which results in association ambiguity, radar frames where an object reflects only one radar point, and extremely low visibility environments that degrade the accuracy of image-based object detection. These corner cases will be addressed in subsequent work to enhance the performance of the presented methodology.

REFERENCES

[1] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, ''A survey of deep learning techniques for autonomous driving,'' *J. Field Robot.*, vol. 37, no. 3, pp. 362–386, Apr. 2020.

[2] L.-H. Wen and K.-H. Jo, ''Deep learning-based perception systems for autonomous driving: A comprehensive survey,'' *Neurocomputing*, vol. 489, pp. 255–270, Jun. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231222003113

[3] B. B. Elallid, N. Benamar, A. S. Hafid, T. Rachidi, and N. Mrani, ''A comprehensive survey on the application of deep and reinforcement learning approaches in autonomous driving,'' *J. King Saud Univ., Comput. Inf. Sci.*, vol. 34, no. 9, pp. 7366–7390, 2022, doi: 10.1016/j.jksuci.2022.03.013.

[4] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, ''Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions,'' *J. Big Data*, vol. 8, no. 1, pp. 1–74, Mar. 2021.

[5] S. Busch, C. Koetsier, J. Axmann, and C. Brenner, ''LUMPI: The Leibniz University multi-perspective intersection dataset,'' in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2022, pp. 1127–1134.

[6] T. Matuszka, I. Barton, A. Butykai, P. Hajas, D. Kiss, D. Kovàcs, S. Kunsàgi-Màté, P. Lengyel, G. Németh, L. Peto, D. Ribli, D. Szeghy, S. Vajna, and B. V. Varga, ''aiMotive dataset: A multimodal dataset for robust autonomous driving with long-range perception,'' in *Proc. Int. Conf. Learn. Represent.*, 2023, pp. 1–28. [Online]. Available: https://openreview.net/forum?id=LW3bRLlY-SA

[7] S. T. Isele, M. P. Schilling, and F. E. Klein, ''Annotating automotive radar efficiently: Semantic radar labeling framework (SeRaLF),'' in *Proc. Conf. Neural Inf. Process. Syst.*, 2020, pp. 1–6.

[8] S. T. Isele, M. P. Schilling, F. E. Klein, S. Saralajew, and J. M. Zoellner, ''Radar artifact labeling framework (RALF): Method for plausible radar detections in datasets,'' 2020, *arXiv:2012.01993*.

[9] N. Scheiner, N. Appenrodt, J. Dickmann, and B. Sick, ''Automated ground truth estimation of vulnerable road users in automotive radar data using GNSS,'' in *IEEE MTT-S Int. Microw. Symp. Dig.*, Apr. 2019, pp. 1–5.

[10] M. Meyer and G. Kuschk, ''Automotive radar dataset for deep learning based 3D object detection,'' in *Proc. 16th Eur. Radar Conf. (EuRAD)*, Oct. 2019, pp. 129–132.

[11] F. Engels, P. Heidenreich, M. Wintermantel, L. Stäcker, M. Al Kadi, and A. M. Zoubir, ''Automotive radar signal processing: Research directions and practical challenges,'' *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 4, pp. 865–878, Jun. 2021.

[12] O. Schumann, M. Hahn, N. Scheiner, F. Weishaupt, J. F. Tilly, J. Dickmann, and C. Wöhler, ''RadarScenes: A real-world radar point cloud data set for automotive applications,'' 2021, *arXiv:2104.02493*.

[13] A. Palffy, E. Pool, S. Baratam, J. F. P. Kooij, and D. M. Gavrila, ''Multi-class road user detection with 3+1D radar in the view-of-delft dataset,'' *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 4961–4968, Apr. 2022.

[14] J. Rebut, A. Ouaknine, W. Malik, and P. Pérez, ''Raw high-definition radar for multi-task learning,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17000–17009.

[15] L. Zheng, Z. Ma, X. Zhu, B. Tan, S. Li, K. Long, W. Sun, S. Chen, L. Zhang, M. Wan, L. Huang, and J. Bai, ''TJ4DRadSet: A 4D radar dataset for autonomous driving,'' in *Proc. IEEE 25th Int. Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2022, pp. 493–498.

[16] S. Agrawal, R. Song, A. Kohli, A. Korb, M. Andre, E. Holzinger, and G. Elger, ''Concept of smart infrastructure for connected vehicle assist and traffic flow optimization,'' in *Proc. 8th Int. Conf. Vehicle Technol. Intell. Transp. Syst.*, 2022, pp. 360–367.

[17] S. Agrawal, R. Song, K. Doycheva, A. Knoll, and G. Elger, ''Intelligent roadside infrastructure for connected mobility,'' in *Smart Cities, Green Technologies, and Intelligent Transport Systems*, C. Klein, M. Jarke, J. Ploeg, M. Helfert, K. Berns, and O. Gusikhin, Eds. Cham, Switzerland: Springer, 2023, pp. 134–157.

[18] C. Creß, Z. Bing, and A. C. Knoll, ''Intelligent transportation systems using external infrastructure: A literature survey,'' 2021, *arXiv:2112.05615*.

[19] S. Kohnert, J. Stähler, R. Stolle, and F. Geissler, ''Cooperative RADAR sensors for the digital test field A9 (KoRA9)—Algorithmic recap and lessons learned,'' in *Proc. Kleinheubach Conf.*, Sep. 2021, pp. 1–4.

[20] A. Krämmer, C. Schöller, D. Gulati, V. Lakshminarasimhan, F. Kurz, D. Rosenbaum, C. Lenz, and A. Knoll, ''Providentia—A large-scale sensor system for the assistance of autonomous vehicles and its evaluation,'' 2019, *arXiv:1906.06789*.

[21] S. Agrawal and G. Elger, ''Concept of infrastructure based environment perception for IN2Lab test field for automated driving,'' in *Proc. IEEE Int. Smart Cities Conf. (ISC)*, Sep. 2021, pp. 1–4.

[22] T. Zhang and P. J. Jin, ''Roadside LiDAR vehicle detection and tracking using range and intensity background subtraction,'' *J. Adv. Transp.*, vol. 2022, pp. 1–14, Apr. 2022.

[23] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "NuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11618–11628.

[24] J. Déziel, P. Merriaux, F. Tremblay, D. Lessard, D. Plourde, J. Stanguennec, P. Goulet, and P. Olivier, "PixSet: An opportunity for 3D computer vision to go beyond point clouds with a full-waveform LiDAR dataset," in *Proc. IEEE Int. Intell. Transp. Syst. Conf. (ITSC)*, Sep. 2021, pp. 2987–2993.

[25] K. Bansal, K. Rungta, S. Zhu, and D. Bharadia, "Pointillism: Accurate 3D bounding box estimation with multi-radars," in *Proc. 18th Conf. Embedded Networked Sensor Syst.*, Nov. 2020, pp. 340–353. [Online]. Available: https://api.semanticscholar.org/CorpusID:227154816

[26] M. Bijelic, T. Gruber, F. Mannan, F. Kraus, W. Ritter, K. Dietmayer, and F. Heide, "Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11679–11689.

[27] D. Yongqiang, W. Dengjiang, C. Gang, M. Bing, G. Xijia, W. Yajun, L. Jianchao, F. Yanming, and L. Juanjuan, "BAAI-VANJEE roadside dataset: Towards the connected automated vehicle highway technologies in challenging environments of China," 2021, *arXiv:2105.14370*.

[28] H. Wang, X. Zhang, Z. Li, J. Li, K. Wang, Z. Lei, and R. Haibing, "IPS300+: A challenging multi-modal data sets for intersection perception system," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2022, pp. 2539–2545. [Online]. Available: https://api.semanticscholar.org/CorpusID:235359104

[29] H. Yu, Y. Luo, M. Shu, Y. Huo, Z. Yang, Y. Shi, Z. Guo, H. Li, X. Hu, J. Yuan, and Z. Nie, "DAIR-V2X: A large-scale dataset for vehicle-infrastructure cooperative 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 21329–21338.

[30] C. Creß, W. Zimmer, L. Strand, M. Fortkord, S. Dai, V. Lakshminarasimhan, and A. Knoll, "A9-dataset: Multi-sensor infrastructure-based dataset for mobility research," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2022, pp. 965–970.

[31] W. Zimmer, C. Creß, H. Tung Nguyen, and A. C. Knoll, "A9 intersection dataset: All you need for urban 3D camera-LiDAR roadside perception," 2023, *arXiv:2306.09266*.

[32] E. Strigel, D. Meissner, F. Seeliger, B. Wilking, and K. Dietmayer, "The Ko-PER intersection laserscanner and video dataset," in *Proc. 17th Int. IEEE Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2014, pp. 1900–1901.

[33] X. Ye, M. Shu, H. Li, Y. Shi, Y. Li, G. Wang, X. Tan, and E. Ding, "Rope3D: The roadside perception dataset for autonomous driving and monocular 3D object detection task," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 21309–21318.

[34] A. Sengupta, A. Yoshizawa, and S. Cao, "Automatic radar-camera dataset generation for sensor-fusion applications," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 2875–2882, Apr. 2022.

[35] Y. Wang, Z. Jiang, X. Gao, J.-N. Hwang, G. Xing, and H. Liu, "RODNet: Radar object detection using cross-modal supervision," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 504–513.

[36] B. Major, D. Fontijne, A. Ansari, R. T. Sukhavasi, R. Gowaikar, M. Hamilton, S. Lee, S. Grzechnik, and S. Subramanian, "Vehicle detection with automotive radar using deep learning on range-azimuth-Doppler tensors," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 924–932.

[37] A. Ouaknine, A. Newson, J. Rebut, F. Tupin, and P. Pérez, "CARRADA dataset: Camera and automotive radar with range-angle-Doppler annotations," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 5068–5075.

[38] M. Sheeny, E. De Pellegrin, S. Mukherjee, A. Ahrabian, S. Wang, and A. Wallace, "RADIATE: A radar dataset for automotive perception in bad weather," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 1–7.

[39] A. Zhang, F. E. Nowruzi, and R. Laganiere, "RADDet: Range-azimuth-Doppler based radar object detection for dynamic road users," in *Proc. 18th Conf. Robots Vis. (CRV)*, May 2021, pp. 95–102.

[40] D.-H. Paek, S.-H. Kong, and K. T. Wijaya, "K-radar: 4D radar object detection for autonomous driving in various weather conditions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds. New Orleans, LA, USA: Curran Associates, 2022, pp. 3819–3829. [Online]. Available: https://proceedings.neurips.cc/paperfiles/paper/2022/file/185fdf627eaae2abab36205dcd19b817-Paper-DatasetsandBenchmarks.pdf

[41] S. Agrawal, S. Bhanderi, K. Doycheva, and G. Elger, "Static multitarget-based autocalibration of RGB cameras, 3-D radar, and 3-D LiDAR sensors," *IEEE Sensors J.*, vol. 23, no. 18, pp. 21493–21505, Sep. 2023.

[42] Z. Zang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, Nov. 2000.

[43] CE Services. (2023). *ARS548 RDI 3D/4D Long Range Radar Datasheet*. [Online]. Available: https://conti-engineering.com/wp-content/uploads/2023/01/RadarSensorsARS548RDI.pdf

[44] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[45] S. Agrawal, S. Bhanderi, S. Amanagi, K. Doycheva, and G. Elger, "Instance segmentation and detection of children to safeguard vulnerable traffic user by infrastructure," in *Proc. 9th Int. Conf. Vehicle Technol. Intell. Transp. Syst.*, 2023, pp. 206–214.

[46] PS Foundation. (2022). *Labelimg Python Tool*. [Online]. Available: https://pypi.org/project/labelImg/

[47] PAG. (2023). *What is Gdpr, the Eus New Data Protection Law?* [Online]. Available: https://gdpr.eu/what-is-gdpr/

[48] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 77–85.

SHIVA AGRAWAL received the B.Eng. degree from the Vidyavardhini's College of Engineering and Technology, Mumbai University, India, in 2012, and the M.Sc. degree in electrical engineering and information technology from Hochschule Darmstadt, Germany, in 2017. He is currently pursuing the Ph.D. degree in sensor perception with AI with the Institute of Innovative Mobility, Technische Hochschule Ingolstadt (THI), Germany. After that, he worked in the industry for four years in ADAS sensor development for AUDI AG at VAIVA GmbH (previously known as Automotive Safety Technologies GmbH), Ingolstadt. He is a Scientific Research Associate with the Institute of Innovative Mobility, THI.

SAVANKUMAR BHANDERI received the Diploma degree in mechanical engineering from the Dr. S. & S. S. Ghandhy College of Engineering & Technology, Surat, India, in 2016, and the B.Eng. degree in mechanical engineering from the Shree Swami Atmanand Saraswati Institute of Technology, Gujarat Technological University, India, in 2019. He is currently pursuing the M.Eng. degree in electrical engineering and information technology with Technische Hochschule Ingolstadt (THI), Germany. He is a part-time Scientific Research Associate with the Institute of Innovative Mobility, THI. His research interests include machine learning and environment perception using multi-sensor systems.

GORDON ELGER received the Diploma degree in physics and the Ph.D. degree from the Free University of Berlin, Germany, in 1994 and 1998, respectively. He then worked with Fraunhofer IZM, Hymite, Electrolux, and Royal Philips in the field of microelectronics packaging. Since 2013, he has been a Professor of electronics and its manufacturing technologies with Ingolstadt University of Applied Sciences (Technische Hochschule Ingolstadt, THI), Ingolstadt. Since 2019, he has been the Head of the Applied Research Center "Connected Mobility and Infrastructure," Fraunhofer IVI, Ingolstadt. His research interests include microelectronic packaging, reliability, and lifetime prognosis of electronic systems, sensor technologies, and sensor data fusion.