IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 January 2023, accepted 7 March 2023, date of publication 17 March 2023, date of current version 3 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3258623

TOPICAL REVIEW

Intelligent Data Handling in Current and Next-Generation Automated Vehicle Development–A Review

MATEUSZ KOMORKIEWICZ[1], (Senior Member, IEEE), **ALVIN CHIN**[2], (Senior Member, IEEE), **PAWEŁ SKRUCH**[3], (Senior Member, IEEE), AND **MARCIN SZELEST**[3], (Senior Member, IEEE)
[1]Aptiv, 30-707 Kraków, Poland
[2]RXO, Charlotte, NC 28277, USA
[3]Department of Automatic Control and Robotics, AGH University of Science and Technology, 30-059 Kraków, Poland

Corresponding author: Mateusz Komorkiewicz (mateusz.komorkiewicz@aptiv.com)

ABSTRACT The higher the vehicle autonomy level is, the more the problems that need to be solved by engineers designing its sensors and control systems. Apart from the mechanical, electrical, algorithmic and validation challenges, such projects require to efficiently gather, store and handle the very large amounts of data logged from the vehicles. This is because the key functionalities of the control system in automated vehicles are based on data that is collected from a variety of sources, processed and analyzed to generate actuation signals. The article goes through the challenges, opportunities as well as solutions associated with data logging, collection, storage, annotation, reprocessing and evaluation. It highlights that effective and efficient data handling is an essential element for obtaining the required performance, reliability, safety and quality capabilities that would allow the mass production of automated vehicles. In terms of the safety of passengers and other road users, the behavior of a mass-produced automated vehicle must be predictable and more reliable than that of an ordinary driver. The multitude of variants that test sequences must cover is practically endless. Therefore, shortening the time-to-market to an acceptable length is only possible if you use the right methods of working with data.

INDEX TERMS Automated vehicles, automotive sensors, big data, smart data logging, system validation.

I. INTRODUCTION

Automated vehicles combine a variety of sensors to perceive their surroundings, including cameras, radars, lidars, GPS, ultrasonic sensors and others. These sensors interpret sensory information to identify navigation paths, avoid obstacles and read relevant markers, like road lanes and signs. Sensors are the base for vehicle perception and without good information about the car's surroundings it will not be possible to achieve higher levels of automation. This can be compared to a human driver with a visual impairment driving a car without glasses. In such a situation, their confidence level of surroundings interpretation is reduced and does not allow them to steer the car safely. In order to improve the performance of the vehicle perception system, car manufacturers and designers increase either the number of sensors mounted on the vehicle or their resolution. In both cases, the amount of raw data generated by the sensory system is drastically rising. Looking at both the modern and future concepts of vehicle system architecture [65], [66], there is a straightforward conclusion that big data is transforming the automotive industry. The key factor that will decide if and when cars with higher level of automation will enter into serial production is how we are able to process this data and analyze it in real time.

The vehicle system architecture for lower levels of autonomy is usually based on separate intelligent sensors, such as radars or cameras. This is because the final decision is rather simple: either to warn the driver or execute lateral

The associate editor coordinating the review of this manuscript and approving it for publication was Divanilson Rodrigo Campelo.

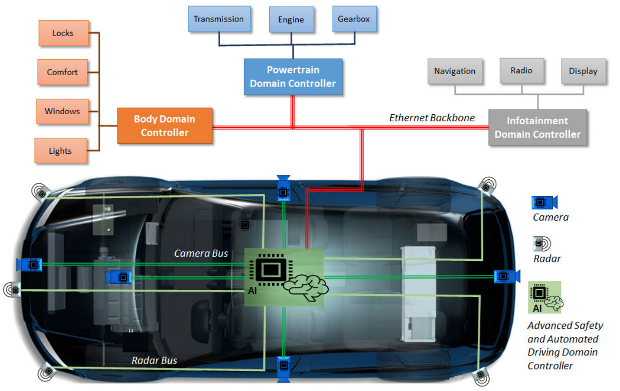

FIGURE 1. New concept of vehicle architecture with Advanced Safety and Automated Driving Domain Controller. The controller integrates unique Artificial Intelligence (AI) functionality allowing to leverage information at the machine level in real time.

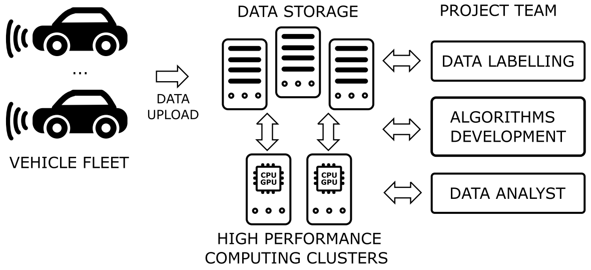

FIGURE 2. Data flow diagram in the development process of automated driving systems.

TABLE 1. Most common automotive sensors and their average data rates [2], [40], [43].

Sensor family	Detailed specification	Frame rate per second	Data rate MB/s
Radars	Detections (CAN)	20	0.03
	Detections (ETH)	20	0.2
	Data Cube (ETH)	20	3
Lidars	32 beams	10	4.7
	64 beams	10	7.4
	128 beams	10	29.3
Cameras	1280x960 - 1.2 MP	30	73.7
	1920x1080 - Full HD	30	124
	3840x2160 - 4K	30	498

and/or longitudinal steering actions within a limited scope of driving scenarios. This decision is based mainly on the analysis of the zone in front of the vehicle. Fig. 1 describes a future concept of vehicle system architecture for advanced safety and automated driving functionalities. In this architecture, the multi-domain controllers [68] (body, powertrain, infotainment, advanced safety and automated driving) are connected to other vehicle systems via Ethernet as a backbone communication network. Sensors can either provide raw and unprocessed data to these units or have some intelligence embedded to allow data pre-processing.

Multi-domain controllers emerged on the market due to several challenges, which are hard to solve by intelligent sensor architecture [45], [68]. The most important are the increased computational complexity of fusion algorithms, which are needed for perception of the environment around the vehicle (360 degrees) and advanced path planning. It is also much easier to prepare heat dissipation mechanisms (e.g. water cooling) for one unit [41].

The development process of both smart sensors and domain controllers requires the cross competency effort of mechanical, thermal, electrical, software, algorithm, verification, system and manufacturing engineers. For function development and verification, data is collected during vehicle test drives in order to create a dataset that can resemble real-world situations as closely as possible. The data flow in a typical automotive project related to automated function development is presented in Fig. 2. First of all, a virtual simulation or real-world vehicle fleet is required to collect data, which can later be used by the project team. The data needs to be physically transferred from the vehicles to the data storage. At this point, it can be accessed by the developers, verification engineers and HPCC (High Performance Computing Clusters). As simple as it might seem, every icon on the diagram means that there is a huge amount of data and along with that comes a long list of technical issues [63].

As the number of different road scenarios is infinite, one of the challenges is how to choose representative data and a second is how much of this data should be collected [46]. Calculation of the maximum rate of data that should be logged leads to another issue related to the design of data loggers that would be able to log all data in a vehicle. The next problem is related to the transfer of the data to the data centers. Besides these problems, others such as data annotation, data reprocessing and data evaluation become real challenges for the development of vehicle control systems intended for series production. The following sections provide more insights on current challenges related to these problems, as well as possible directions for solving them.

II. DATA LOGGING

Depending on the autonomy level [34] that has to be implemented in a given project, different sensor setups must be used (e.g. cameras, radars [47]) in the car [32]. Along with the production sensor set, additional reference sensors (e.g. lidar) must be used that are at least orders of magnitude more precise in order to make testing and validation effective. Depending on the sensor's type, different data is generated. The following sections contain, in summary, the characteristics of data logged by the most common automotive sensors.

A. RADARS

Radars analyze the reflections of transmitted electromagnetic waves in order to extract detections from different objects [48]. A single detection carries information not only about the position of a reflection point but also a number of other interesting parameters, such as target speed (from the doppler effect) etc. [28]. Radar data can be logged on several levels [62]. The first is raw ADC (analog-digital converter) values captured from the antennas. The second is data cube (DC) representation obtained after applying the Fast Fourier transform (FFT) to raw data. The third is detection – point cloud of extracted reflections. The final is the object level – aggregated information about detected objects, such as cars, pedestrians etc. In general, the higher the abstraction level is, the lower the amount of data generated by radar. As for now, the most common level of radar data logging is the detection level. Two interfaces are usually used for this: CAN (Controller Area Network) [25], where the number of detections is limited to 32-64 entries, and Ethernet with the number of detections usually above 128. The estimated data rates for DC, CAN and Ethernet streams [43] are presented in Table 1.

B. CAMERAS

In the production vehicle, the images from the camera are captured and transferred directly to an embedded processing system which outputs only high-level information about detected objects [20]. This means that the video itself is not available in the production vehicle (except parking systems, e.g. rear-view camera). For project development purposes (algorithm training dataset, electronic control unit (ECU) validation, in-the-loop verification), the raw video must be recorded as well. Camera video stream data size is a function of image resolution, dynamic range bit depth and frame rate [40]. Every pixel has a single value that is usually encoded in 12 to 16 bits (high dynamic range imagers are used). The frame rate ranges usually from 20 to 40 Hz. The three most common image resolutions used in automotive Advanced Driver Assistance Systems (ADAS) are listed in Table 1. The data is usually logged using different proprietary LVDS (Low-Voltage Differential Signaling) solutions [77].

C. LIDARS

Lidar technology [44] is still not mature enough (both in cost per unit as well as lifetime durability) to be mounted on a mass scale onto vehicles. There are, however, some examples of successful projects using solid state lidar technology [29]. Nevertheless, in most cases, lidars are used as a reference sensor to enable the precise labeling and positioning of objects around vehicle that is logging the data. The most popular due to the quality vs. price factor are still the rotating devices [54]. The amount of data generated by lidar is related to angular resolution, the number of vertical beams and rotation rate [2]. It is often described in points per second (PPS) [50]. For 32 beam lidars, the number of points received per second is about 1.3M reaching about 6M for high-end 128 beam units. For every point, the distance and reflectivity is returned (usually encoded on 3 or 4 bytes). The estimated data bandwidth is presented in Table 1. Ethernet is most commonly used for data transmission.

TABLE 2. Estimated daily data rates depending on the assumed number of sensors in different project types.

Project (SAE Level)	Cameras (resolution)	Radars /Lidars	MB/s	TB per one day of logging (16h)
Level 1 radar	0	7/1	8.76	0.48
Level 1 cameras	3 x Full HD	0/1	363	19.96
Level 2+	8 x 1.2 MP	0/0	563	30.90
Level 3+	1 x 4K 5 x Full HD 8 x 1.2 MP	4/1	1639	90.01

D. OTHER SENSORS AND DATA SOURCES

There are also other sensors, such as ultrasonic [76] or precise Global Navigation Satellite System (GNSS) receivers [37], but their data rate is usually in the range of a few kilobytes per second; thus, they can be excluded from logger data bandwidth requirement computation. The vehicle, however, generates a lot of inter ECU traffic (body, engine, steering) that often needs to be recorded as well. In particular, information about the car state, such as speed or power mode etc., is necessary in order to re-simulate the car environment. The bandwidth is usually in the range of a few megabytes per second and requires the logging of multiple LIN (Local Interconnect Network), CAN, FlexRay and BroadR-Reach channels [57].

E. FULL SETUP

The data bandwidths presented in Table 1 are for single sensors only. Depending on the project complexity (autonomy level), multiple sensor configurations are commonly used. In Table 2, a short comparison is given between different Society of Automotive Engineers (SAE) level vehicles [34] for a common sensor setup. The numbers for level 2+ are estimated based on sensor setup from [11]. For level 3+, the estimation was done by the authors based on marketing materials of companies such as Lucid Motors (32 sensors, 14 cameras) [8] and 5th-generation Waymo Driver (29 cameras) [10].

In order to carry out project development successfully, all data needs to be logged. It is important to understand the complexity of such an endeavor.

The first problem is related to the amount of data that is generated by the vehicle. In most cases, the data (including camera images) captured and saved during logging campaigns can only be compressed by lossless codecs to ensure that no data is lost [27], [58]. This is necessary because only raw unmodified data can be used for training new algorithms and computing performance indicators. This requirement implies that the recorded data size is large.

Based on the numbers presented in Table 2, it is clear that the requirements towards automotive data loggers are quite high. The logger must also support a wide range of different input interfaces, such as LVDS, Ethernet, CAN, FlexRay [57] etc. When it comes to data storage requirements, it is worth mentioning that often special SSD (solid-state drive) arrays must be used to ensure that no data is lost during recording.

The next issue is that with so many sensors, variable frame rates and multiple data busses, accurate timestamping becomes crucial to enable the proper re-use of logged information. This is not a trivial undertaking due to the different principles of operations, latency in data processing and transmission, as well as other factors. Therefore time stamping of data packets at the logger side is not accurate and is considered not good enough for most applications (e.g. data fusion). The solution to this problem is to distribute the master clock to all sensors in order to synchronize them and enable time stamping at the sensor end. But this requires special hardware and software solutions as for example lidars are synchronized with GPS, radars with CAN master clock frames, cameras with trigger signals and there is no simple way to synchronize all sensors.

Another problem which needs to be tackled in order to enable data usage is related to the movement of the vehicle used for logging. For example, a rotating lidar on the roof rack of a moving car will change its origin with every firing sequence. With 10Hz rotation and 60km/h vehicle speed, the displacement will be more than 1.5m for a single scan. Thus, the points sampled at the beginning of the scan cannot be simply merged with points at the end of the scan without compensating for the car movement.

The final challenge related to the logging setup is its power requirements. With every additional reference sensor and logging device, the entire system's power requirements increase as well. In passenger cars, as well as in light commercial vehicles with 12V electrical installations, it is quite easy to exceed the standard serial production alternator capabilities because of the additional equipment. This means the vehicle electric power generation and distribution system might require major rework.

One solution to the problems mentioned in this chapter is the shift towards Ethernet (1Gb/10Gb) universal loggers. With special gateways, switches and multiple 10Gb Ethernet ports in one logger, it is possible to cope with the problem of logging huge amounts of data from so many sources. Unfortunately, the problem of synchronization is still not entirely solved by the Ethernet-based logging setup (especially due to the variety of sensors). As such, it could possibly be an interesting field for new standards definition and normalization. It should be emphasized that standard Ethernet is used for test vehicles in the signal logging setup. In the serial production of vehicles, the BroadR-Reach standard is used, which differs from Ethernet in its physical layer [13].

F. SMART DATA LOGGING

The problem with enormous amount of data generated during logging campaigns can be solved by an additional system that would analyse the recorded data either in the logging vehicle and/or in the data center [24], [74]. The general idea is presented in Fig. 3. In the vehicle it requires additional computer or embedded device that is able to analyze (preferably with Artificial Intelligence algorithms) the content of the data that is recorded. It can either use the original sensors mounted in the vehicle (from the system that is developed) or additional sensor (e.g. lidar, camera, radar etc.). Once the system knows what is in the logs it can be used to delete (in the vehicle) the portion of logs which does not contain new (e.g. 1h drive on an motorway with only one vehicle in front) or not relevant (e.g. dark road) sequences. The algorithm that is making the decision has to take into consideration not to remove all the negative samples (random negative logs can be stored) in order to preserve the correct positive to negative logs ratio. In literature, this approach is called event-based data collection [72]. The vehicle is equipped with a detection system and before data collection starts, a list of events that triggers storage is defined. The list contains signals from many sources: speed, yaw rate, acceleration, number of visible objects, class of objects, interactions between objects, radar detections, type of environment, weather conditions etc. Since some of the events are computationally complex and data recording triggered by a specific event needs to cover information a few seconds before the event occurs (this problem is discussed in the section about re-processing), the logging system needs to be equipped with a cyclic buffer that is large and fast enough to handle and store all sensor data without any frame drops. Unfortunately, even in the case of perfect logging equipment, it is impossible to eliminate entirely the risk of missing important events not previously defined on the list of events.

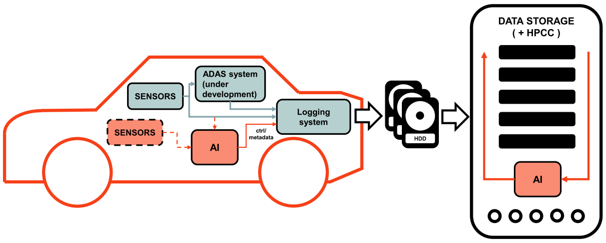

FIGURE 3. Intelligent logging concept.

Online log analysis in the vehicle can save a lot of logger disk space and lower the amount of data that needs to be transferred to the data storage (for more details on data storage see Section III-C). The same analysis can be performed again in the data storage center. All the logs can be analyzed with the Machine Learning (ML) algorithm to extract the number of objects and other interesting metadata. It can be used both to tag the recordings and to make a decision if some logs can be removed.

III. DATA COLLECTION

It would seem that smart data logging solves the problem of collecting and analyzing large amounts of data - unfortunately not in all cases. One such case is re-simulation, described later in this article. Introducing the re-simulation to the appropriate

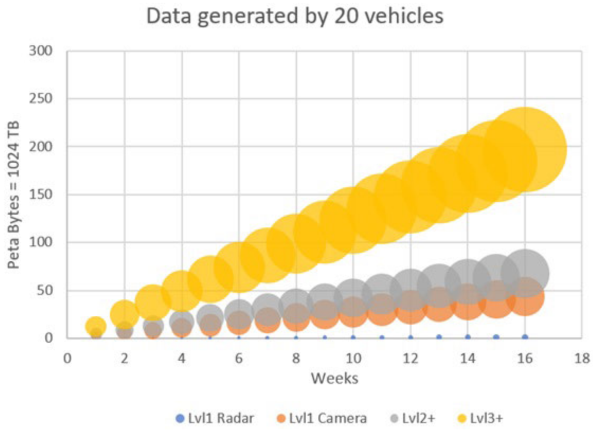

FIGURE 4. Estimated data size growth for various projects.

initial condition takes time, so practically speaking, the re-sim must be fed with data at least several seconds before the event appears. In the case of logs with a lot of events (and these are the most interesting from an engineering point of view), event-based logging turns into continuous logging. A data collection campaign aims to collect data from representative road scenarios. As the number of different road scenarios is practically infinite, selecting a subset of those that can be considered as representative is a really challenging task.

A. REAL WORLD USER PROFILE (RWUP) DATA

In current projects intended to develop system components for series production, the data collection process is performed in actual road traffic. This is called RWUP (Real-World User Profile) scenarios [59]. At the initial stages of the project, this involves a ride in so-called open-loop settings. In this case, the system is installed in a car and its outputs are logged but are not set on the actuators. Closed-loop tests are conducted in a controlled environment on a dedicated test field area. As the data has to be collected over a short time (usually a few months), in the test campaign, a car fleet of a dozen up to several dozens of cars is involved. Each car is driven by a driver and a technician who monitors whether the logging system works correctly. The cars collect the data in two or three 8-hour shifts. Based on the amount of data generated by one vehicle per day (16h of logging), which is presented in Table 2, it is possible to estimate the numbers for the vehicle fleet (the reasonable fleet size of 20 vehicles is assumed). Within four months, the amount of generated data ranges from a few (for radar-only projects) to hundreds of petabytes (for more complex projects – Level 3 and above) which is presented in Fig. 4.

B. VIRTUAL WORLD USER PROFILE (VWUP) DATA

Although increasingly more advanced tools are available that allow the generation of virtual test scenarios [33], still the ratio of using such data in the development process of advanced safety and automated driving systems is small. Car manufacturers still require having data collected from real test drives. The main reason for this might be the fact that synthetic data does not allow the proper evaluation of the performance of perception systems. The need to change, however, is noticeable [31], [64].

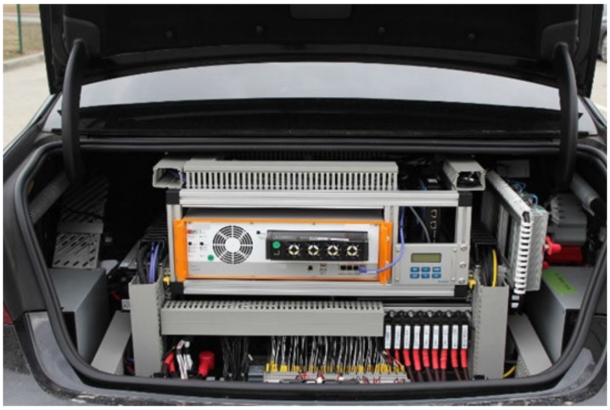

FIGURE 5. Data logging setup in a vehicle trunk.

C. DATA STORAGE

A cursory analysis of the requirements for data logging and collection gives a picture of the complexity of data handling [17]. Under an optimistic assumption to log all sensor data with total bandwidth far above the data rate on a single hard drive and to deliver and upload all data (hundreds of thousands of kilometers of driving) to a data center on time, the total size of collected data will be around 30 PB (according to calculations shown on Fig. 4). As it is difficult to imagine how much 30 PB of data is, we can check how many typical data carriers should be used to save the required amount of data. If we use single-layer DVD discs (capacity 4.7 GB) to store RWUP, we would need 6.6 millions of them. If we use 2 TB hard disks, we would need about 15k units. Another challenge is to transfer such a large amount of data to a computing center (private storage or cloud). Handling 15,000 hard drives, setting up a dedicated fiber optic connection or using dedicated media for copying data to a cloud is a huge and costly logistical challenge.

D. TRUNK VOLUME

Test cars drive in many countries, on several continents and collect data. Due to the high data rate from the logger, the physical dimensions of the vehicle become a challenge. It turns out that, especially on longer routes, the size of the trunk is of great importance (see Fig. 5). Storage, plus a logger, plus two drivers can easily exceed the permissible load value. Bad weight distribution can seriously affect sensor calibration and an overloaded vehicle has different motion physics, which can make the logged data worthless. It is also necessary to find a smart way to unload a car from recorded data: copy to a cloud, data truck etc. [73].

E. DATA AVAILABILITY

Another important factor is the time that passes from when the test is performed until the data is ready for analysis. The driver of the vehicle does not have the time and competence to verify that a recording is correct (correct scenario, exhaustive set of signals, correct calibration of sensors etc.). It should be noted that a single missing signal may necessitate the repetition of the entire data collection. In order to avoid situations like this, some preventive measures should be implemented, such as:

- Artificial Intelligence based real-time anomaly detection in a logger (to avoid missing signals) [18];
- on-the-fly analysis of metadata (to maintain planned RWUP distribution);
- possibility of uploading 1% of logs to a data center on the same day they are recorded (for manual analysis of the recording by an experienced engineer) [49].

F. SAFETY BACKUP

As the car fleet is driving worldwide, an average driving time for each car may be 16h per day. In the remaining 8h, all data from the car should be copied and sent to a data center. This means that the copy rate must be at least twice as fast as the recording rate [9]. In order to minimize the risk of data loss, two copies of the data should be made. This again doubles the expected copy data rate.

G. DATA EXCHANGE MEDIUM

Reading and copying data from the vehicle in order to release the vehicle's buffer and enable the continuation of data logging, requires the use of appropriate data carriers. Remembering that the data destination is the Data Center, an appropriate data exchange medium needs to be used [3], [12]. Direct upload to a cloud is not an option due to limited bandwidth and lack of network in the vehicle. Dedicated data exchange boxes can be used obtained from Cloud providers, as well as regular hard disk drives (HDD) but:

- they should be fast enough to copy all data in less than 8h;
- they should be large enough to accommodate all data;
- they should be delivered every day to different destinations and picked up after 8h;
- there should be enough of them to cover a fleet of 20 vehicles.

Here is a rough estimation: 2 boxes per vehicle every day, 1 day to deliver boxes to the data hub, 1 day for copy, 1 day to deliver empty boxes, at least 8 storage boxes per car (160 boxes for a 20-car fleet). On top of this, we should also make provisions for a backup copy (which simply doubles storage needs).

H. DATA MANAGEMENT SYSTEM

The correct logging of data, synchronization and delivery to a data center does not guarantee their optimal use. As the fleet of cars are driving worldwide, hundreds of petabytes of recorded data are received daily and are delivered safely and reliably to our data center. In order to ensure the continuity of the flow of data boxes, we must be sure that every day we upload everything and empty the source carrier. This is not a trivial task. In order to upload 1 *PB* of data (250 HDDs of 4 *TB* each) weekly, we need to copy the data in parallel (the copy of a single 4 *TB* drive lasts over 10h). Proper data indexing is a critical issue, as reviewing the collected data repeatedly is completely unjustified economically. To that effect, we need to use a Data Management System (DMS) [23], [46]. DMS consists of stored data and corresponding metadata with information about speed, yaw rate, country of origin, license plate of vehicle etc. DMS should be able to add further metadata at a later time. A very valuable source of metadata, available later in the project, can be the output data from the re-simulation. This data contains information about detections (pedestrians, vehicles), as well as weather conditions and road profile estimations. Due to the amount of collected data and limitations of local area network bandwidth, it is not possible to work with raw data. Raw data is necessary for resimulation (see Section V for more details) but, especially for manual analysis, it is more convenient to work with data extract. The development team needs to work with metadata plus extracted and compressed streams that can be queried according to that metadata.

I. DATA RETENTION

Another significant challenge is data retention. From a data perspective, the project consists of the following phases: sensors and logger development, data collection, software and hardware development of processing unit, start of series production and maintenance of the product. By default, data should be available for analysis for a period of ten years from the end of production. This means that the total duration of data storage can be up to 15 years. There are two basic ways to store data: Hot Storage, where the data is immediately available, and Cold Storage, where the data is stored externally, usually on magnetic tapes, and one must wait several days for it to be copied back to the hot storage [35]. The cost of storing petabytes of data for such a long period of time on Hot Storage (high-speed and high-bandwidth file system) is significant. Therefore, it is necessary to consider storing the data externally on magnetic tapes (Cold Storage). In the data collection phase, around 20% of total space required to store RWUP data is needed to be on Hot Storage (for all new files, which automatically, after a few weeks, go to Cold Storage) and 80% on Cold Storage. In the system development phase, 100% of storage is required to be on Hot Storage. When series production starts, all data can be moved to Cold Storage and, in addition, 20% of Hot Storage should be available for the maintenance phase.

J. DATA PRIVACY

Another challenge with data storage and handling is related to data privacy acts, which have been established around the globe (see also [61]). In Europe, GDPR (General Data Protection Regulation) [6] is in effect; China proposed PDPL (Personal Data Protection Law) [5] and in California, USA there is CCPA (California Consumer Privacy Act) [4]. Unfortunately, data collected by autonomous vehicles (i.e. cameras) contains personal identification data (license plates, faces etc.), which requires special handling. Due to the fact that a single project requires data collection in many countries with different data privacy regulations, it makes it extremely hard for the project team and function owners to be able to access and reprocess data in a single location. There are approaches with video anonymization (blurring sensitive areas like in popular navigation applications) but this heavily impacts the results of video reprocessing.

K. ACCESSING DATA

With petabytes of data in local data centers or remote cloud storage, one of the biggest issues is to ensure that the project team will be able to use data efficiently. There are several activities that require data access: labeling, log re-simulation, algorithm development, data analysis and functional analysis. This means that the data needs to be accessed on a daily basis (not entire datasets though) by more than one hundred people in different geographical locations, as well as HPCC (most of the dataset) in one or several locations. Unfortunately, the truth is that there is no network that could withstand working with raw logs on a daily basis. One of the solutions to this problem is for engineering teams to work with compressed logs (rather than raw logs), whenever possible (for example for labeling and data analysis), and ensure a high bandwidth link between the data center and HPCC (re-simulation requires raw uncompressed logs). With so much data, it becomes crucial to pick the logs which are most interesting and contain objects and situations from as many perception modules as possible [60]. This approach allows to label and re-simulate a smaller number of logs but test all required functions and modules at the same time.

L. DATA DISTRIBUTION SPECIFICATION

Datasets that are used in the automotive field are specific in terms of their size, properties and diversity [38]. The description of the datasets should take the form of metadata characterizing scenes and scenarios. The metadata should reflect not only static but also dynamic aspects in the recorded video sequences. The data collection process is carried out in the majority of the countries where car manufacturers intend to sell their cars. A graphical illustration of the data collection campaign can be found in Fig. 6. Table 3 illustrates an example of data collection specification for validation of an acceptable false alarm rate in Automatic Emergency Braking (AEB) system according to the Safety of the intended functionality norm [56]. Although this specification contains a set of quantified parameters, it is not clear how the recommended distribution of the values of these parameters can guarantee the unambiguous and comprehensive description of the dataset.

M. GRID BASED CONCEPT FOR DATA COLLECTION SPECIFICATION

A possible approach for the formal description of the automotive datasets is based on the grid concept [42]. According to this concept, the sensor's field of view is represented by a 2D or 3D grid, whose size can be set based on the sensor accuracy (see Fig. 7). As a first step, a set of the measurable properties characterizing the scene (or scenario), which are important from a developed function perspective, should be identified. Then, every property can be considered as a random variable, which is defined on a space containing grid elements. In addition, a probability distribution function should be assigned to every random variable. The distribution function gives the cumulative probability of the occurrence of a specific event on the grid elements. During data collection, the histogram or heatmap of occupied grid cells by recorded sequences can be compared with the expected distribution in order to verify the representativeness of the dataset. Illustrating this concept, the heatmap of grid occupancy by tracks provided by a 360-degree surround radar-based perception has been calculated and is shown in Fig. 8 [42].

The presented grid approach falls into the category of scenario-based approaches [52] where certain situations experienced while driving are assigned to the list of predefined scenarios [22], [70]. The scenarios may represent maneuvers of the objects on the scene and interactions between them [69], [71]. However, the grid approach does

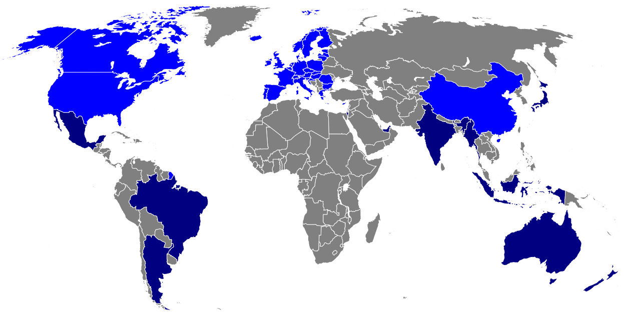

FIGURE 6. Illustration of country distribution in data collection campaign. Areas marked in blue represent the primary scope (around 80%) of test drives and those marked in dark blue the secondary scope (around 20%), which have been calculated based on new car sales in 2019 [1].

TABLE 3. Example of data collection specification (2019 – ISO/PAS 21448:2019 Road vehicles – Safety of the intended functionality).

Time of Day		
Type		**Percentage**
Day		50 %
Night		35 %
Dusk		15 %
Vehicle Speed		
Speed [mi/h]	**Speed [km/h]**	**Percentage**
0 – 25	0 – 40	60 %
26 – 50	41 – 80	40 %
$>$ 50	$>$ 80	0 %
Type		**Percentage**
Dry/Clear sky		65 %
Rain		10 %
Fog		5 %
Snow		5 %
Overcast		10 %
Heavy rain		5 %

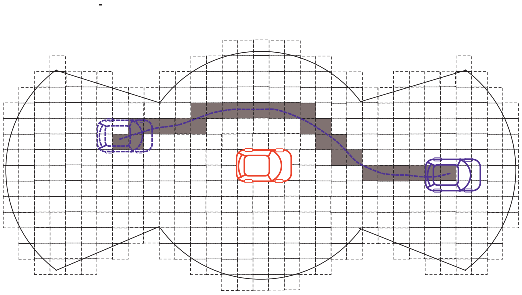

FIGURE 7. A sensor's field of view is represented by a 2D grid. Object tracks are selected as the property characterizing the scenarios. The picture presents grid occupancy by one track.

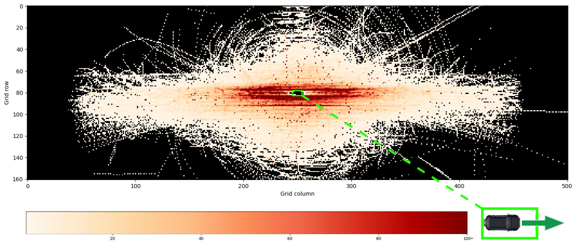

FIGURE 8. Heatmap representing grid occupancy by tracks in a radar system with 360-degree field of view. The size of a grid element is 1*m* × 1*m*.

not require a catalog of the scenarios and by adding to the grid an additional dimension representing the speed, the dynamics of the scenario can be included to this concept. Another application of the scenario-based approach is described in [24] where predefined events of interest represent anomalies in video data. As the scenarios are characterised by some parameters, so a set of the performance indices calculated per scenario can be also used to evaluate the collected data [23].

IV. DATA LABELING (ANNOTATION)

Manual labeling is a very important yet tedious task, which requires a special approach. This can range from simple 2D bounding box drawing on camera images and 3D bounding box marking on lidar point clouds to multi-point line marking, image segment annotation [67] and full-point cloud segmentation labeling [75]. Drawing a 2D bounding box rectangle around objects like cars or pedestrians on camera images might seem well defined. There are, however, situations when it is hard to decide, even for an algorithm developer, if and how the object should be marked. For example, if a vehicle or pedestrian is 20% occluded by another vehicle, should it be marked only in a visible area or as a whole object? It is not obvious how to annotate vehicles with a roof rack, a pedestrian with headgear or one whose torso or legs are only visible.

Each task requires different effort, which is usually expressed in time of labeling per one second of recorded data. Of course, this might vary from frame to frame (e.g. different number of objects on the motorway versus in the city) but it can be somehow averaged over a long labeling campaign. Even if marking a 2D bounding box in a camera image takes 1 second (choose object class, double-click over object etc.) and on average there are 30 objects on a scene (vehicles, traffic signs, pedestrians), 30 seconds are needed to label one image frame. With 30 fps cameras, this results in 900 seconds of labeling for 1 second of recorded data, which is 15 minutes. With a recording fleet of 20 cars, recording for 16h per day for 1 month (20C-16C-30C-3600), it can be computed that labeling all data would take 8.6 million hours, which is almost 1000 years!

The above result brings into question the issue of labeling scalability. It is very hard for a human to focus on such a task for more than a few hours. Due to human exhaustion and carelessness, manual labeling is plagued with errors [30]. This is why labeling needs to be distributed between hundreds of labelers. Another challenge is the verification of labeling work, which also requires manpower. Finally, the more labelers are working on one project, the bigger the number of data annotation differences and errors that will be introduced to the dataset. In general, it is not so simple to scale the labeling effort and keep the quality of reference data on a par.

A lot of effort nowadays is put into preparing Artificial Intelligence (AI) based auto-labeling algorithms, which can either replace or augment the manual labeling effort [21], [53]. This is important because even a small reduction in the time required to label one second of logged data results in huge savings in project time and money. The problem with this approach, however, is that it is hard to use AI to train another AI and expect that the trained algorithm will be better than the teacher. Of course, some techniques can be used to improve the auto-labeling results (e.g. log forward/backward pass) [36] but, in general, the expectation of vehicle manufacturers is that every auto-labeled frame needs to be verified and corrected by a human labeler. The final question is whether applying corrections makes labeling faster than labeling from scratch.

Instead of labeling the entire log, one can only mark the most interesting parts based on detected events. Event-based labeling consists in saving the state of the sensors and their detections from the moment of the event occurrence for a specified time forward, to the memory, only in case one of the defined sequences of events has occurred. This approach significantly reduces the workload, but it should be remembered that the event does not come out of nowhere and, in a way, it requires prior labeling. In addition, resimulation, Software in the Loop (SiL) or Hardware in the Loop (HiL) [16] require entering the initial state (or rather close to the initial state, as it is not possible to modify the values of local variables in the tested device), which means that a sufficient amount of data (frames) should be injected into the device before the event; labels assigned prior to this event cannot be used to count Key Performance Indicators (KPIs).

V. DATA USAGE–RESIMULATION

Vehicular systems that process data captured by the vehicle's sensors (radar, camera, lidar, ultrasonic) and provide signals

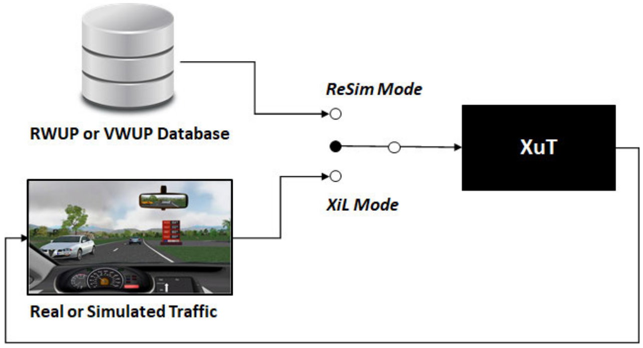

FIGURE 9. Data reprocessing strategies in ReSim and XiL (MiL, SiL, PiL, HiL, ViL) modes for XuT (Model under Test, Software under Test, Device under Test, Vehicle under Test).

to drive the vehicle, are very complex. They utilize multiple different algorithms and neural networks, and implement a variety of perception and planning modules (e.g. Traffic Sign Recognition, Lane Detection, Path Planning etc.). The effort needed to test such a complex system is significant. Assuming that the design team provides new software releases every two weeks, the entire testing procedure should be completed between two consecutive project sprints. Tests should cover the full distribution of the expected road scenarios and weather conditions, in which the final product is going to operate. This corresponds to hundreds of thousands of km of distance that should be covered between the releases. In order to reduce the need for repetition of test drives for each software release, a common approach is to record data only once and reuse it later virtually. The technique to use RWUP for incoming software releases is called re-simulation (ReSim) [59].

RWUP data is injected to ReSim (see Fig. 9), and ReSim acts as a model of recently released software. After closing the testing loop (by virtual scenario or human interaction), ReSim can be performed on different abstraction levels: MiL (Model in the Loop), SiL (Software in the Loop), PiL (Processor in the Loop), HiL (Hardware in the Loop) and ViL (Vehicle in the Loop). In general, the lower the description level, the higher computation effort is needed to process the same distance. HiL and ViL need to process in real time, while MiL, SiL and PiL do not. Due to this, MiL, SiL and PiL are easy to scale, with relatively low cost, on HPCC. Please note that the overall test procedure consists of a combination of the above-mentioned abstraction levels. Some features have to be tested in HiL and the vehicle (with the highest cost and lowest scalability) but the majority may be covered by MiL, SiL and PiL.

It is worth mentioning that RWUP is only used for open loop ReSim. Therefore, with this data we can only compare detections between a real vehicle and a software model. In order to analyze the impact of detections on the behavior of a vehicle, we need to close the simulation loop. To that effect, we need to replace RWUP with a virtually created environment that has the ability of being modified in response to the vehicle's behavior.

Nowadays, most of the testing tasks related to the verification of advanced safety and automated driving systems, is executed with SiL and HiL [19]. SiL integrates third-party libraries and takes control of feeding them with relevant data (RWUP or virtual environment). The SiL binary is encapsulated into containers (Docker or Singularity). Many containers may be initiated in parallel on HPCC to re-simulate separate logs. This feature makes it possible to re-simulate thousands of kilometers in reasonable time. Unfortunately, target software needs to be exported to SiL libraries and what causes that behavior of SiL is slightly different than the behavior of HiL and ViL – not all HW features are used or properly modeled in SiL libraries. Moreover, the development of the SiL environment is permanently delayed to the target software release, as additional time to create libraries and integration, plus testing are needed.

SoC (System on Chip) is much bigger and much more complex than processors of HPCC [14]. As a result, SiL for recent SoC is significantly slower than real time. In the near future, this difference will become large enough that using SiL will become pointless. HiL uses final hardware flashed with the recent software version. This solution is the most accurate and closest to the behavior of the vehicle. HiL, by default, works in real time, which is its big advantage, since it is as close as possible to a real vehicle. Since HiL works in real time, input data has to be delivered continuously without any delays or interruptions. This is a big challenge for scalability because of storage bandwidth where RWUP data is kept. HiL is also much more expensive than SiL. A single HiL device may cost a few hundred thousands of dollars. Both SiL and HiL are created for certain projects and their re-use of hardware/software for other projects is strongly limited. In summary, SiL is a good approach to simulate large amounts of data in reasonable time but in the next few years it will become obsolete due to the complexity difference between HPCC processors and dedicated SoC. HiL is more akin to the real world but is very expensive, and the scalability of HiL devices is limited by storage bandwidth and the virtual environment. Both SiL and HiL are designed for certain projects, so their re-use is not possible.

VI. DATA EVALUATION

The ability to verify the performance of the ADAS or Autonomous Driving (AD) system under development is crucial from both the development team and vehicle manufacturer perspectives. This is why, in every project, evaluation metrics called Key Performance Indicators (KPIs) are defined [55].

Although it might seem that it should be rather straightforward to use mathematical formulas and statistics to compute KPIs and obtain information about system performance, the reality is much more complex.

First of all, in order to compute KPIs the ground truth must be available. This, in most cases, requires the manual labeling of reference sensor data. In real projects, only a small fraction of recorded data is used for evaluation. It will simply take too long to label all data, and without proper labeling the metrics cannot be computed.

This generates a problem of how to choose the right subset for labeling. It turns out that selecting data by random sampling is not the best solution. This is because when a logging campaign is planned and executed, the differentiation in scenarios is made purely based on kilometers of recording with a given set of parameters, such as road type, weather, lighting conditions etc. Thus, with random sampling, the system will be verified against the most common road conditions (few cars on the motorway / traffic jam in a city etc.) but not against the most problematic corner cases.

As it was mentioned in Section IV even manual labeling is not ideal. The most commonly used evaluation metrics are based on intersection over union (IOU) of 2D or 3D bounding boxes (marking objects) [39]. The labelling inaccuracy (e.g. one or two pixel bounding box position shift) affect the bounding box size and the final IOU value.

It is also worth mentioning that well-established academic techniques for measuring the accuracy of objects detection and tracking (e.g. [15]) are in fact evaluating only the statistical performance of perception ADAS/AD system [51]. They still do not answer the question of whether the system will be able to use this knowledge to react properly (e.g. perform emergency braking). This is because the statistics does not take into consideration that missing an object in front of the vehicle in 1% of cases can be deadly, whereas missing it 10% of cases on the rear does not really affect vehicle safety.

Thus, the evaluation should be done not only on the sensor perception level but also on the system level. For example, authorities in California, USA require the reporting of AD systems' disengagement per miles driven to be prepared for all AD cars tested in the state [7].

But the more KPI metrics are introduced, the harder it is to use them to define the final system performance and to use it to compare different software/hardware releases. Usually, one checklist is prepared with pass/fail results for all KPIs (with an acceptance threshold set by the customer). If every point has a pass status, this is binary information indicating that everything is fine. Unfortunately, during product development that is not true. In such a case, it is not so simple to decide whether a new update is beneficial when it improves one metric but exacerbates another.

Some of these problems are addressed by the new Safety of the Intended Functionality standard (ISO/PAS 21448:2019 Road vehicles – Safety of the intended functionality) [26]. SOTIF proposes how to define a risk and validation strategy in order to evaluate end function safety. It proposes to split scenarios into four different categories: known safe, known unsafe, unknown unsafe, unknown safe. The goal of the project development team is to reduce the number of unsafe scenarios at the end of the project, as well as to define a statistical approach to ensure that the number of unknown unsafe scenarios is acceptably low compared to the known safe and unsafe scenarios.

VII. CONCLUSION

The finding of this study suggested multiple challenges in data processing research area that are needed to be solved for automated vehicles intended for series production. The development and verification of such complex systems require large datasets, where properties such as volume, velocity, variety, variability and veracity can be formally defined. Formal description of the datasets shall include qualitative and quantitative measures characterizing both static and dynamics aspects of the recorded test scenarios. Having such measures, the completeness and adequacy of the datasets can be evaluated.

Data is the only way to prove the acceptable level of safety, reliability, robustness, performance, security and other system properties required for the vehicles. The exponential growth of raw data generated by the sensors' suite dictates that only representative test scenarios may be collected in an already pre-processed format. In this context, event based (or selective) data recording seems to be more appropriate comparing to continuous data logging. The research activities should be then focused on intelligent data loggers. Improved lossless data or video codecs can also result in less data to be stored. In addition, as the data collection campaign is an extremely time-consuming task in the system development process, the use of data from one project in others should be a common practice.

Data annotation using automatic tools is absolutely necessary. However, the quality of such tools must be comparable to the manual annotation performed by human-experts. Moreover, annotation of single scans without taking into account preceding and following scans will not give full picture on what is happening on the scene. Thus the process should evaluate towards video classification and characterization.

The ratio of using VWUP versus RWUP must be increased and it should not be set arbitrarily but should follow on from the distribution of a set of properties in the dataset. In order to increase usage of the virtual simulation data, good models of the sensors should be developed. This can be considered as the key element missing in the race for automated driving.

The last but not least issue is that the whole development and verification process of the automotive control system relies on supervised approaches. Consequently, huge effort is required to get sufficient amount of ground truth information (reference data). Therefore, the application of self-supervised or unsupervised techniques could revolutionize the way how the automated vehicle are designed and verified.

REFERENCES

[1] *Annual Figures: New Registrations*. Accessed: Dec. 1, 2022. [Online]. Available: https://www.vda.de/en/news/facts-and-figures/annual-figures/new-registrations

[2] *Application Note VLP-16: Packet Structure & Timing Definition*. Accessed: Dec. 1, 2022. [Online]. Available: https://velodynelidar.com/wp-content/uploads/2019/09/63-9276-Rev-C-VLP-16-Application-Note-Packet-Structure-Timing-Definition.pdf

[3] *AWS Snowball*, Amazon, Seattle, WA, USA, Feb. 27, 2022.

[4] *California Consumer Privacy Act*, State California Dept. Justice, Sacramento, CA, USA, Feb. 27, 2022.

[5] *China NPC Law (Chinese)*, Nat. People's Congr. People's Republic China, Beijing, China, Feb. 27, 2022.

[6] *Complete Guide to GDPR Compliance*, Eur. Union, Brussels, Belgium, Feb. 27, 2022.

[7] *Disengagement Reports*. Accessed: Mar. 14, 2022. [Online]. Available: www.dmv.ca.gov/portal/vehicle-industry-services/autonomous-vehicles/disengagement-reports

[8] *The Full Reveal: Dreamdrive*. Accessed: Feb. 21, 2022. [Online]. Available: www.lucidmotors.com/stories/grand-reveal-dreamdrive

[9] *How Long Should My Backup Take in 2022?* Accessed: Dec. 1, 2022. [Online]. Available: https://www.cloudwards.net/how-long-should-my-backup-take/

[10] *Introducing the 5th-Generation Waymo Driver: Informed by Experience, Designed for Scale, Engineered to Tackle More Environments*. Accessed: Feb. 21, 2022. [Online]. Available: blog.waymo.com/2020/03/introducing-5th-generation-waymo-driver.html

[11] *Tesla's Hardware Retrofits for Model 3*. Accessed: Feb. 8, 2022. [Online]. Available: www.eetasia.com/teslas-hardware-retrofits-for-model-3/

[12] *What is Azure Data Box?* Microsoft, Redmond, WA, USA, Feb. 27, 2022.

[13] *Broad-R-Reach Physical Level Transceiver Specification For Automotive Applications*, V 3.0, Broadcom Corporation, Bengaluru, Karnataka, May 2014.

[14] A. Ahmad, "Automotive semiconductor industry—Trends, safety and security challenges," in *Proc. 8th Int. Conf. Rel., Infocom Technol. Optim. (Trends Future Directions) (ICRITO)*, Jun. 2020, pp. 1373–1377.

[15] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The CLEAR MOT metrics," *EURASIP J. Image Video Process.*, vol. 2008, pp. 1–10, Dec. 2008.

[16] C. Brogle, C. Zhang, K. L. Lim, and T. Braunl, "Hardware-in-the-loop autonomous driving simulation without real-time constraints," *IEEE Trans. Intell. Vehicles*, vol. 4, no. 3, pp. 375–384, Sep. 2019.

[17] H. Caesar, V. Bankiti, H. A. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuScenes: A multimodal dataset for autonomous driving," 2019, *arXiv:1903.11027*.

[18] F. Caetano, P. Carvalho, and J. Cardoso, "Deep anomaly detection for in-vehicle monitoring—An application-oriented review," *Appl. Sci.*, vol. 12, no. 19, p. 10011, Oct. 2022.

[19] S. Chen, Y. Chen, S. Zhang, and N. Zheng, "A novel integrated simulation and testing platform for self-driving cars with hardware in the loop," *IEEE Trans. Intell. Vehicles*, vol. 4, no. 3, pp. 425–436, Sep. 2019.

[20] S. Dabral, S. Kamath, V. Appia, M. Mody, B. Zhang, and U. Batur, "Trends in camera based automotive driver assistance systems (ADAS)," in *Proc. IEEE 57th Int. Midwest Symp. Circuits Syst. (MWSCAS)*, Aug. 2014, pp. 1110–1115.

[21] M. Dimitrievski, I. Shopovska, D. Van Hamme, P. Veelaert, and W. Philips, "Automatic labeling of vulnerable road users in multi-sensor data," in *Proc. IEEE Int. Intell. Transp. Syst. Conf. (ITSC)*, Sep. 2021, pp. 2623–2630.

[22] H. Elrofai, J.-P. Paardekooper, E. D. Gelder, S. H. Kalisvaart, and O. O. D. Camp, "StreetWise: Scenario-based safety validation of connected automated driving," TNO, The Hague, The Netherlands, Tech. Rep., 2018.

[23] F. Farahani and F. Rezaei, "Implementing a scalable data management system for collected data by smart meters," in *Proc. 26th Int. Comput. Conf., Comput. Soc. Iran (CSICC)*, Mar. 2021, pp. 1–5.

[24] R. Feng, Y. Yao, and E. Atkins, "Smart black box 2.0: Efficient high-bandwidth driving data collection based on video anomalies," *Algorithms*, vol. 14, no. 2, p. 57, Feb. 2021.

[25] International Organization for Standardization. *Road Vehicles—Controller Area Network (CAN)—Part 1: Data Link Layer and Physical Signalling*, Standard ISO 11898-1:2015, 2019.

[26] International Organization for Standardization. *Road Vehicles—Safety of the Intended Functionality*, Standard ISO/PAS 21448:2019, 2019.

[27] J. Forster, X. Jiang, A. Terzis, and A. Rothermel, "Evaluation of compression algorithms for automotive stereo matching," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2012, pp. 1017–1022.

[28] M. E. Gadringer, F. M. Maier, H. Schreiber, V. P. Makkapati, A. Gruber, M. Vorderderfler, D. Amschl, S. Metzner, H. Pflugl, W. Bösch, M. Horn, and M. Paulweber, "Radar target stimulation for automotive applications," *IET Radar, Sonar Navigat.*, vol. 12, no. 10, pp. 1096–1103, Oct. 2018.

[29] C. Galle, J. Amelung, T. Dallmann, and S. Brueggenwirth, "Vehicle environment recognition for safe autonomous driving: Research focus on solid-state LiDAR and RADAR," in *Proc. Automot. Meets Electron., 11th Gmm-Symp.*, 2020, pp. 1–3.

[30] N. Ghahreman and A. B. Dastjerdi, "Semi-automatic labeling of training data sets in text classification," *Comput. Inf. Sci.*, vol. 4, no. 6, p. 48, Oct. 2011.

[31] M. Hahner, D. Dai, C. Sakaridis, J.-N. Zaech, and L. V. Gool, "Semantic understanding of foggy scenes with purely synthetic data," in *Proc. IEEE Intell. Transp. Syst. Conf. (ITSC)*, Oct. 2019, pp. 3675–3681.

[32] M. Hartstern, V. Rack, M. Kaboli, and W. Stork, "Simulation-based evaluation of automotive sensor setups for environmental perception in early development stages," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Oct. 2020, pp. 858–864.

[33] S. Hasirlioglu and A. Riener, "A general approach for simulating rain effects on sensor data in real and virtual environments," *IEEE Trans. Intell. Vehicles*, vol. 5, no. 3, pp. 426–438, Sep. 2020.

[34] SAE International. *Taxonomy and Definitions for Terms Related to Driving Automation Systems for on-Road Motor Vehicles*, Standard J3016:202104, Warrendale, PA, USA, 2021.

[35] R. Irie, S. Murata, Y.-F. Hsu, and M. Matsuoka, "A novel automated tiered storage architecture for achieving both cost saving and QoE," in *Proc. IEEE 8th Int. Symp. Cloud Service Comput. (SC2)*, Nov. 2018, pp. 32–40.

[36] L. Jacobs, A. Kodumuri, J. James, S. Park, and Y. Kim, "Multiperspective automotive labeling," in *Proc. IEEE Int. Parallel Distrib. Process. Symp. Workshops (IPDPSW)*, May 2020, pp. 927–936.

[37] N. Joubert, T. G. Reid, and F. Noble, "Developments in modern GNSS and its impact on autonomous vehicle architectures, in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Oct. 2020, pp. 2029–2036.

[38] Y. Kang, H. Yin, and C. Berger, "Test your self-driving algorithm: An overview of publicly available driving datasets and virtual testing environments," *IEEE Trans. Intell. Vehicles*, vol. 4, no. 2, pp. 171–185, Jun. 2019.

[39] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, and J. Zhang, "Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 319–336, Feb. 2009.

[40] I. Kolak, Z. Lukac, M. Knezic, and S. Koncar, "Realization of automotive video data acquisition system for usage in evolution of autonomous vehicles," in *Proc. Zooming Innov. Consum. Technol. Conf. (ZINC)*, May 2020, pp. 160–164.

[41] J. Korta, P. Skruch, and K. Holon, "Reliability of automotive multidomain controllers: Advancements in electronics cooling technologies," *IEEE Veh. Technol. Mag.*, vol. 16, no. 2, pp. 86–94, Jun. 2021.

[42] P. Kowalczyk, M. Komorkiewicz, P. Skruch, and M. Szelest, "Efficient characterization method for big automotive datasets used for perception system development and verification," *IEEE Access*, vol. 10, pp. 12629–12643, 2022.

[43] D. Kraus, H. Ivanov, and E. Leitgeb, "Approach for an optical network design for autonomous vehicles," in *Proc. 21st Int. Conf. Transparent Opt. Netw. (ICTON)*, Jul. 2019, pp. 1–6.

[44] Y. Li and J. Ibanez-Guzman, "LiDAR for autonomous driving: The principles, challenges, and trends for automotive LiDAR and perception systems," *IEEE Signal Process. Mag.*, vol. 37, no. 4, pp. 50–61, Jul. 2020.

[45] L. Liu, S. Lu, R. Zhong, B. Wu, Y. Yao, Q. Zhang, and W. Shi, "Computing systems for autonomous driving: State of the art and challenges," *IEEE Internet Things J.*, vol. 8, no. 8, pp. 6469–6486, Apr. 2021.

[46] A. Luckow, K. Kennedy, F. Manhardt, E. Djerekarov, B. Vorster, and A. Apon, "Automotive big data: Applications, workloads and infrastructures," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Oct. 2015, pp. 1201–1210.

[47] E. Marti, M. A. de Miguel, F. Garcia, and J. Perez, "A review of sensor technologies for perception in automated driving," *IEEE Intell. Transp. Syst. Mag.*, vol. 11, no. 4, pp. 94–108, Winter 2019.

[48] S. M. Patole, M. Torlak, D. Wang, and M. Ali, "Automotive radars: A review of signal processing techniques," *IEEE Signal Process. Mag.*, vol. 34, no. 2, pp. 22–35, Mar. 2017.

[49] C. Prehofer and S. Mehmood, "Big data architectures for vehicle data analysis," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2020, pp. 3404–3412.

[50] Aravind Ratnam. *PPS: Finally, a Consistent Metric to Gauge LiDAR Performance*. Accessed: Nov. 29, 2022. [Online]. Available: https://sensephotonics.com/pps-finally-a-consistent-metric-to-gauge-lidar-performance/

[51] F. Reway, W. Huber, and E. P. Ribeiro, "Test methodology for vision-based ADAS algorithms with an automotive camera-in-the-loop," in *Proc. IEEE Int. Conf. Veh. Electron. Saf. (ICVES)*, Sep. 2018, pp. 1–7.

[52] S. Riedmaier, T. Ponn, D. Ludwig, B. Schick, and F. Diermeyer, "Survey on scenario-based safety assessment of automated vehicles," *IEEE Access*, vol. 8, pp. 87456–87477, 2020.

[53] S. Roychowdhury and L. S. Muppirisetty, "Fast proposals for image and video annotation using modified echo state networks," in *Proc. 17th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2018, pp. 1225–1230.

[54] S. Royo and M. Ballesta-Garcia, "An overview of LiDAR imaging systems for autonomous vehicles," *Appl. Sci.*, vol. 9, no. 19, p. 4093, Sep. 2019.

[55] M. N. Sharath and B. Mehran, "A literature review of performance metrics of automated driving systems for on-road vehicles," *Frontiers Future Transp.*, vol. 2, p. 28, Nov. 2021.

[56] D. Shen, Q. Yi, L. Li, R. Tian, S. Chien, Y. Chen, and R. Sherony, "Test scenarios development and data collection methods for the evaluation of vehicle road departure prevention systems," *IEEE Trans. Intell. Vehicles*, vol. 4, no. 3, pp. 337–352, Sep. 2019.

[57] Y.-H. Sheu, T.-H. Fu, C.-M. Ku, J.-J. Yang, and Y.-Y. Hsu, "The design of FlexRay/CAN/LIN protocol analyzer," in *Proc. Int. Conf. Electric Inf. Control Eng.*, Apr. 2011, pp. 5466–5469.

[58] C. Shuai, S. Li, and H. Liu, "Comparison of compression algorithms on vehicle communications system," in *Proc. IEEE Adv. Inf. Technol., Electron. Autom. Control Conf. (IAEAC)*, Dec. 2015, pp. 91–95.

[59] P. Skruch, R. Dlugosz, K. Kogut, P. Markiewicz, D. Sasin, and M. Rozewicz, "The simulation strategy and its realization in the development process of active safety and advanced driver assistance systems," SAE, Warrendale, PA, USA, Tech. Rep., 2015-01-1401, 2015.

[60] K. Sung, K.-W. Min, J. Choi, and B.-C. Kim, "A formal and quantifiable log analysis framework for test driving of autonomous vehicles," *Sensors*, vol. 20, no. 5, p. 1356, Mar. 2020.

[61] D. Suo, J. Siegel, and A. Soley, "Driving data dissemination: The 'Term' governing connected car information," *IEEE Intell. Transp. Syst. Mag.*, vol. 13, no. 1, pp. 20–30, Spring 2021.

[62] P. Swami, A. Jain, P. Goswami, K. Chitnis, A. Dubey, and P. Chaudhari, "High performance automotive radar signal processing on TI's TDA3X platform," in *Proc. IEEE Radar Conf. (RadarConf)*, May 2017, pp. 1317–1320.

[63] J. Tian, A. Chin, and M. Karg, "Digital services in the automotive industry," *IT Prof.*, vol. 18, no. 5, pp. 4–6, Sep. 2016.

[64] J. Tremblay, A. Prakash, D. Acuna, M. Brophy, V. Jampani, C. Anil, T. To, E. Cameracci, S. Boochoon, and S. Birchfield, "Training deep networks with synthetic data: Bridging the reality gap by domain randomization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 1082–10828.

[65] J. P. Trovao, "Trends in automotive electronics [automotive electronics]," *IEEE Veh. Technol. Mag.*, vol. 14, no. 4, pp. 100–109, Dec. 2019.

[66] J. P. Trovao, "Automotive electronics market evolution [automotive electronics]," *IEEE Veh. Technol. Mag.*, vol. 15, no. 1, pp. 107–118, Mar. 2020.

[67] P. Voigtlaender, L. Luo, C. Yuan, Y. Jiang, and B. Leibe, "Reducing the annotation effort for video object segmentation datasets," 2020, *arXiv:2011.01142*.

[68] D. Wang and S. Ganesan, "Automotive domain controller," in *Proc. Int. Conf. Comput. Inf. Technol.*, Sep. 2020, pp. 1–5.

[69] W. Wang and D. Zhao, "Extracting traffic primitives directly from naturalistically logged data for self-driving applications," *IEEE Robot. Autom. Lett.*, vol. 3, no. 2, pp. 1223–1229, Apr. 2018.

[70] H. Weber, J. Bock, J. Klimke, C. Rosener, J. Hiller, R. Krajewski, A. Zlocki, and L. Eckstein, "A framework for definition of logical scenarios for safety assurance of automated driving," *Traffic Injury Prevention*, vol. 20, pp. 65–70, Jun. 2019.

[71] H. Weber, J. Hiller, B. Metz, T. Louw, Y. M. Lee, R. Madigan, N. Merat, E. Lehtonen, H. Sintonen, S. Innamaa, T. Streubel, L. Pipkorn, E. Svanberg, M. Weperen, J. Hogema, A. Bolovinou, M. Junghans, M. Zhang, J. Trullos, and L. Eckstein, "L3pilot deliverable 7.3: Pilot evaluation results," L3Pilot Consortium, Tech. Rep., Oct. 2021.

[72] Z. Xie, M.-L. Shyu, and S.-C. Chen, "Video event detection with combined distance-based and rule-based data mining techniques," in *Proc. IEEE Multimedia Expo Int. Conf.*, Jul. 2007, pp. 2026–2029.

[73] Y. Yan, L. Zhang, Q. Z. Sheng, B. Wang, X. Gao, and Y. Cong, "Dynamic release of big location data based on adaptive sampling and differential privacy," *IEEE Access*, vol. 7, pp. 164962–164974, 2019.

[74] Y. Yao and E. Atkins, "The smart black box: A value-driven high-bandwidth automotive event data recorder," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 3, pp. 1484–1496, Mar. 2021.

[75] L. Yi, B. Gong, and T. Funkhouser, "Complete & label: A domain adaptation approach to semantic segmentation of LiDAR point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15363–15373.

[76] J. Yu, S. E. Li, C. Liu, and B. Cheng, "Dynamical tracking of surrounding objects for road vehicles using linearly-arrayed ultrasonic sensors," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2016, pp. 72–77.

[77] V. Zwillich, M. Wollitzer, T. Wirschem, W. Menzel, and H. Leier, "Signal integrity analysis of a 1.5 Gbit/s LVDS video link," in *Proc. IEEE Int. Symp. Electromagn. Compat.*, Jul. 2007, pp. 1–6.

MATEUSZ KOMORKIEWICZ (Senior Member, IEEE) received the Ph.D. degree (summa cum laude) in automation control and robotics from the Faculty of Electrical Engineering, Automatics, Computer Science and Electronics, AGH University of Science and Technology, Kraków, Poland, in 2010. He is currently a Senior Engineer of computer vision and artificial intelligence with the Advanced Engineering Department, Aptiv. His research interests include deploying AI solutions on SoC devices as well as nonstandard ML usage in the automotive industry.

ALVIN CHIN (Senior Member, IEEE) received the Ph.D. degree in computer science from the University of Toronto, in 2007. Currently, he is the Principal Data Scientist in the Brokerage department at RXO. Previously, he was an AI and Emerging Technology Researcher at the Tech Office Silicon Valley and Senior Machine Learning Researcher at the Tech Office Chicago of the BMW Group. He is also an Adjunct Professor at the DePaul University College of Computing and Digital Media. His research interests are related to deploying AI for business use cases, data science and connected vehicles. He is the Chair of the IEEE Chicago Section.

PAWEŁ SKRUCH (Senior Member, IEEE) received the Ph.D. degree (summa cum laude) in automation control from the Faculty of Electrical Engineering, Automatics, Computer Science and Electronics, AGH University of Science and Technology, Kraków, Poland, in 2005, and the D.Sc. (Habilitation) degree in automatics and robotics from the AGH University of Science and Technology, in 2016. He is currently a Professor of control engineering with the AGH University of Science and Technology and a Manager for new AI/ML applications with the Aptiv Technical Center, Kraków. His current research interests include dynamical systems, autonomous systems, artificial intelligence, machine learning, modeling and simulation, and the applications of control theory to software systems.

MARCIN SZELEST (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the Silesian University of Technology, Gliwice, Poland, in 2014. From 2004 to 2008, he was an Integrated Circuit Designer with Evatronix Company. In 2009, he has decided to move to the automotive industry working with Tier 1 Company Aptiv (formerly, Delphi). He is the coauthor of 14 records of the invention and over 15 scientific publications. His research interests include the development of robust electrical devices, high-performance computing for the resimulation of active safety systems, software-defined radio, and driver and cabin monitoring systems. In 2019, he was one of the founders of the Polish Chapter of IEEE VTS and became the first Chairman.