Vision-Based Semantic Segmentation in Scene Understanding for Autonomous Driving: Recent Achievements, Challenges, and Outlooks

Khan Muhammad, *Senior Member, IEEE*, Tanveer Hussain, *Student Member, IEEE*, Hayat Ullah, *Student Member, IEEE*, Javier Del Ser, *Senior Member, IEEE*, Mahdi Rezaei, *Member, IEEE*, Neeraj Kumar, *Senior Member, IEEE*, Mohammad Hijji, *Member, IEEE*, Paolo Bellavista, *Senior Member, IEEE*, and Victor Hugo C. de Albuquerque, *Senior Member, IEEE*

***Abstract*— Scene understanding plays a crucial role in autonomous driving by utilizing sensory data for contextual information extraction and decision making. Beyond modeling advances, the enabler for vehicles to become aware of their surroundings is the availability of visual sensory data, which expand the vehicular perception and realizes vehicular contextual awareness in real-world environments. Research directions for scene understanding pursued by related studies include person/vehicle detection and segmentation, their transition analysis, lane change, and turns detection, among many others. Unfortunately, these tasks seem insufficient to completely develop fully-autonomous vehicles *i.e.,* achieving level-5 autonomy, travelling just like human-controlled cars. This latter statement is among the conclusions drawn from this review paper: scene understanding for autonomous driving cars using vision sensors still requires significant improvements. With this motivation, this survey defines, analyzes, and reviews the current achievements of the scene understanding research area that mostly rely on computationally complex deep learning models. Furthermore, it covers the generic scene understanding pipeline, investigates the performance reported by the state-of-the-art, informs about the time complexity analysis of avant garde modeling choices, and highlights major triumphs and noted limitations encountered by current research efforts. The survey also includes a comprehensive discussion on the available datasets, and the challenges that, even if lately confronted by researchers, still remain open to date. Finally, our work outlines future research directions to welcome researchers and practitioners to this exciting domain.**

***Index Terms*— Autonomous driving, autonomous vehicles, context prediction, deep learning, scene understanding, semantic segmentation.**

Manuscript received 28 April 2021; revised 21 December 2021 and 18 July 2022; accepted 14 September 2022. Date of publication 6 October 2022; date of current version 5 December 2022. This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 957339. The Associate Editor for this article was S. Wan. *(Corresponding author: Khan Muhammad.)*

Khan Muhammad is with the Visual Analytics for Knowledge Laboratory (VIS2KNOW Laboratory), Department of Applied Artificial Intelligence, School of Convergence, College of Computing and Informatics, Sungkyunkwan University, Seoul 03063, Republic of Korea (e-mail: khan.muhammad@ieee.org).

Tanveer Hussain and Mahdi Rezaei are with the Institute for Transport Studies, University of Leeds, LS2 9JT Leeds, U.K. (e-mail: tanveer445@ieee.org; m.rezaei@leeds.ac.uk).

Hayat Ullah is with the Intelligent Systems, Computer Architecture, Analytics, and Security Laboratory (ISCAAS Laboratory), Department of Computer Science, Kansas State University, Manhattan, KS 66506 USA (e-mail: hayatullah@ieee.org).

Javier Del Ser is with the TECNALIA, Basque Research & Technology Alliance (BRTA), 48160 Derio, Spain, and also with the Department of Communications Engineering, University of the Basque Country (UPV/EHU), 48013 Bilbao, Spain (e-mail: javier.delser@tecnalia.com).

Neeraj Kumar is with the Department of CSED, Thapar University, Patiala, Punjab 147004, India, also with the School of Computer Science, University of Petroleum and Energy Studies, Dehradun, Uttarakhand 248121, India, also with the Department of Electrical and Computer Engineering, Lebanese American University, Beirut 1102 2801, Lebanon, and also with Department of Computing and IT, King Abdul Aziz University, Jeddah 22230, Saudi Arabia (e-mail: nehra04@gmail.com; neeraj.kumar@thapar.edu; neeraj.kumar.in@ieee.org).

Mohammad Hijji is with the Faculty of Computers and Information Technology (FCIT), University of Tabuk, Tabuk 47711, Saudi Arabia (e-mail: m.hijji@ut.edu.sa).

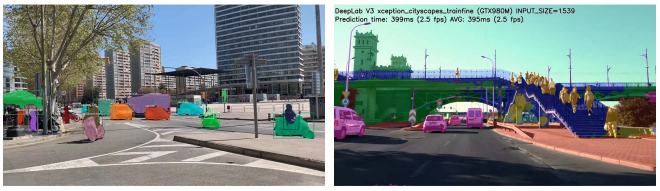
Paolo Bellavista is with DISI, University of Bologna, 40126 Bologna, Italy (e-mail: paolo.bellavista@unibo.it).

Victor Hugo C. de Albuquerque is with the Department of Teleinformatics Engineering, Federal University of Ceará, Fortaleza, Ceará 60455-970, Brazil (e-mail: victor.albuquerque@ieee.org).

Digital Object Identifier 10.1109/TITS.2022.3207665

I. Introduction

AUTONOMOUS Driving (AD) relies on processed information from numerous sensors installed over the vehicle, perceiving the surroundings, helping to understand the traffic scenes and control the movements of the vehicle [1], and hence playing a role of its eyes and ears. These sensors mostly include high resolution cameras, radar, and Light Imaging Detection and Ranging (LiDAR) [2] to classify the objects via feature extraction and to measure the distance to surrounding objects via radio waves and illumination, so as to eventually yield a 3D view of the environment. To avoid collision with on-road obstacles, various types of other sensors have also been deployed for autonomous vehicles, which include infrared, sonar, micro radar, ultrasonic, and short distance sensors. Similarly, vision sensors are used to equip autonomous vehicles with the ability to understand the visuals of surrounding environment, which include road lanes detection, traffic light analysis, road sign detection and recognition, vehicle detection



(a) Instance segmentation (b) Semantic segmentation

Fig. 1. Sample segmented images for an autonomous vehicle, helping in scene parsing: (a) exemplifies instance segmentation, where each object from similar classes is segmented into different color with its own boundary pixels; (b) depicts a semantically segmented image, where objects of similar classes are highlighted in an individual color, without any differentiation.

and tracking, pedestrian detection (both on-road and off-road), and short-term traffic prediction [3]. Visual scene representation and understanding for AD include lanes detection, traffic lights analysis, traffic signs, surrounding pedestrian and cars detection, and many other tasks. Accumulating these information provide more enhanced and safer instructions for automated actions of the vehicle, such as turning manoeuvres, lane changing, or braking [4].

Among the various sources of information gathered for vehicular decision making, vision sensors data [5] are arguably considered as the most reliable ones [6]. Therefore, this research domain has been extensively studied and widely applied in Intelligent Transportation Systems (ITSs) [7], mostly from a machine learning perspective and by resorting to deep Convolutional Neural Networks (CNNs). Deep CNNs embody a special flavor of neural networks with several functional layers suitable to process images by repetitively extracting model features from the input image, towards optimally achieving better representations. Scene understanding from vision data operates likewise, applying a deep CNN over real-time frames to *e.g.* interpret a pedestrian location and its distance from the autonomous car. Beyond this simplified generic computer vision-based scene understanding, complex models proposed nowadays are able to generate multiple labelled outputs (*e.g.* pedestrians and vehicles), as well as their localization.

Scene understanding primarily refers to context extraction from visual data that is based on different features such as shapes of objects, their distance from the vehicle, and many other clues including size of the objects and their approaching speed. A scene analysis can be achieved by accumulating these information and building a complete scenario of the scene around the vehicle, so that vehicular systems can be informed of *e.g.* the presence of humans in front of the car and their distance from the autonomous vehicle. When assessed together, this information helps in actions being taken by the autonomous vehicle, where the distinction among various humans, vehicles, buildings, traffic signs, turns *etc.* is essential for proper decision making, as visualized in Figure 1. Traditionally, these information streams are extracted in isolation using separate computer vision algorithms [11], which are recently replaced by CNNs-based segmentation mechanisms. A segmentation mechanism annotates the boundaries of various types of objects and assigns different colors to each pixel identified to belong to different objects. Pixel-level labeling may refer to semantic or instance segmentation as shown in Figure 1, where instance-level segmentation assigns different colors to each object, even in the same class (*e.g.* vehicles), whereas pixel-level semantic segmentation assigns the same color labels for the same class of objects. Among many traditional segmentation strategies [12], the most widely used category is semantic segmentation using deep CNNs, which partitions an ongoing scene into different meaningful elements such as road, cars, pedestrians, trees, besides other elements present in the vehicular context.

A. Background and Related Works

Semantic segmentation is widely used in AD applications until proper scene understanding [13] demands a clear distinction between two identical objects. For example, surrounding cars pose a similar label in semantic segmentation networks, and convey a clear understanding of the scene for further decision making. However, at some point, AD needs instance-level segmentation to deal with various types of traffic stakeholders and their levels of engagement. Traditionally, there are three representative types of semantic segmentation networks represented in Figure 2: fully convolutional networks (FCNs), deep fully convolutional neural network architecture for semantic pixel-wise segmentation known as (SegNet), and the so-called DeepLab strategy [14], which we briefly revisit next towards arriving at the purpose of this manuscript.

To begin with, the FCNs [8] architecture is structured in encoder-decoder formation to extract deep discriminative features for later instance localization and segmentation task. The encoder part comprises of standard convolutional and down-sampling layers typically used in CNNs for classification problems, where the decoder part used transposed convolutional layer to up-sample the coarse output feature maps from the bottleneck layers of the architecture as shown in Figure 2(a). The up-sampling process can be achieved at coarse and finer level of FCNs *i.e.,* instead of the traditional last layer output, it can be passed through transposed convolution layer(s), that help produce prediction maps of the same size as the input frame. On the other hand, SegNet is built upon a series of deconvolutional layers that transform the extracted features into class score prediction maps as an output with identical frame size as that of the input. A SegNet network comprises two functional modules, where the first extracts features from the input frame using a CNN and the prediction maps with class scores are constructed via a series of transposed convolutions and un-pooling layers in the second module [15], ultimately producing instance-level segmentation results. This kind of segmentation strategy is also known as *encoder-decoder strategy*. Finally, the DeepLab strategy for semantic segmentation utilizes convolutional layers with an up-sampled filter, known as *atrous* convolutions, with bilinear interpolation to obtain prediction maps of identical size as an input frame.

Recently a plethora of new semantic segmentation methods [16], [17], [18], [19], [20], [21], [22], [23], [24] for visual scene understanding has emerged in the literature, eliciting impressive results. For instance, Nesti *et al.* [19] presented a method that evaluates the robustness of semantic segmentation

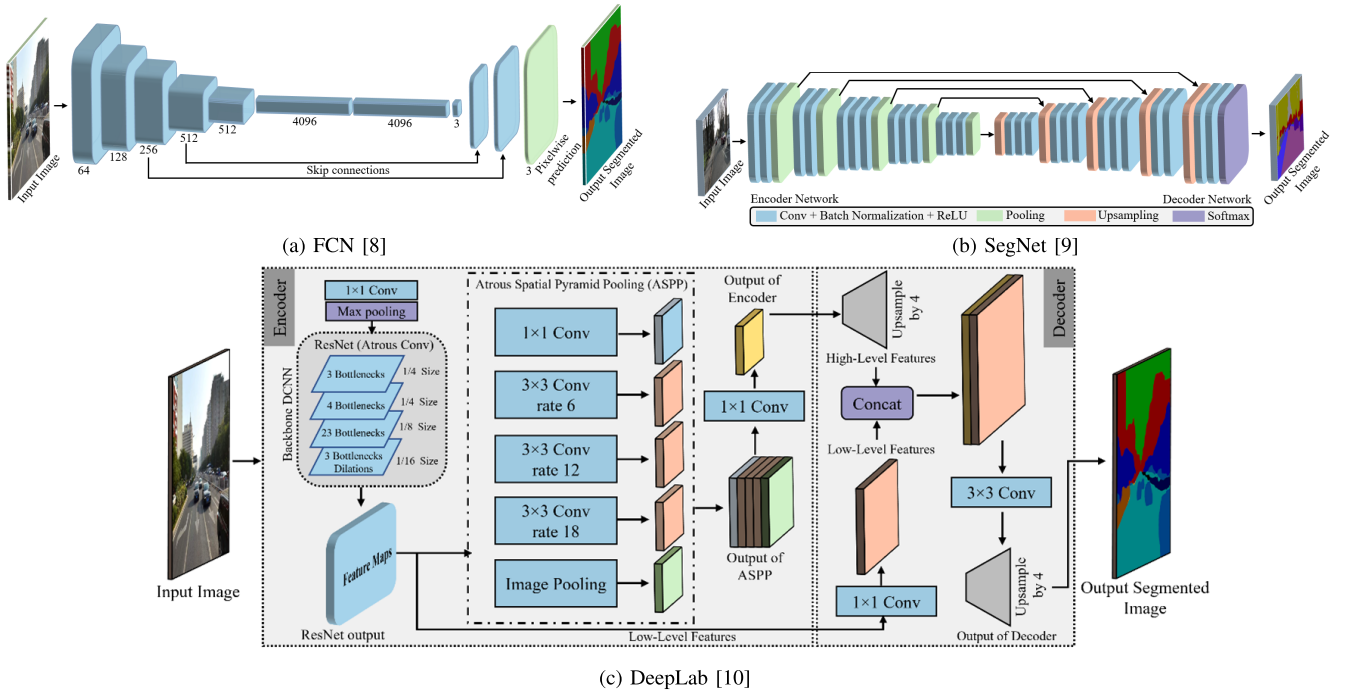


Fig. 2. Various segmentation network architectures, adopted by mainstream research in segmentation as baseline strategies.

approaches for autonomous vehicles. They introduced a novel loss function to analyze the effectiveness of existing semantic segmentation methods against real-world adversarial attacks in autonomous driving environments. Natan *et al.* [21] proposed a compact yet efficient multi-task learning semantic segmentation method to deal with different modes of data. Their method has the ability to perform various tasks in a unified approach that includes depth estimation, semantic segmentation, ranging (LiDAR data) segmentation, and light detection. To analyze the model uncertainty problem for semantic segmentation, Zhao *et al.* [22] presented a pyramid bayesian approach, which evaluates the uncertainity of semantic segmentation model for autonomous driving. They examined the performance of semantic segmentation model (SegNet) by replacing dropout layers with pyramid pooling layer and claimed improvement in their model's performance. Baran *et al.* [24] introduced a unique approach for understanding the road view semantics through onboard Bird's Eye View (BEV) camera visuals. They have analyzed the understating of road scenery in three different perspectives that include image-level, BEV level, and aggregated temporal road scene understanding. Traditionally, neural networks are trained using powerful graphics processing units (GPUs) and huge server computers, whereas inference is performed over embedded systems in self-driving cars. Lately the computational complexity has been reduced significantly by some deep models such as SqueezeNet [25], which achieves AlexNet level accuracy with 50 times less number of parameters. Following SqueezeNet, ENet [26] achieved real-time semantic segmentation over embedded devices. More recently, semantic segmentation achieved significant milestones from the perspective of time complexity, as reported in [25] and [26]. For the better understanding of readers, the graphical overview of the major architectures distribution of the semantic segmentation driven scene understanding literature for autonomous driving is depicted in Figure 2.

B. Challenges and Motivation

Self-driving cars have to react instantly according to the surroundings, where in real-world circumstances there are higher chances to encounter new type of events, putting the car in tangle situation. Furthermore, the inherent uncertainty associated to unknown situation increases the probability of the model to issue erroneous decisions, putting the lives of passengers and other counterparts nearby in danger. The inference of a trained model installed in self-driving cars needs to be dynamic in nature, perceiving real-time decisions, aware of the confidence in its own outputs, learning from new events, and updating the parameters of their model. Similarly, decisions made by self-driving cars are mostly generated by black-box neural models, leaving a manifold of open questions for explainable and accountable decisions made by an autonomous car. Moreover, future location perception of pedestrians and vehicles with truly actionable accuracy is still to be achieved in AD. Similarly, complex driving scene understanding and visual scene perceptions in adverse weather conditions are also open challenges yet to be covered in AD domain. All these challenges are of utmost necessity to see driver-less cars moving safely in urban areas. Unfortunately, despite prior efforts [29], [30], the community lacks a consolidated, unified, single point of reference for ascertaining the current level of maturity of semantic segmentation techniques for vehicular scene understanding.

Considering the aforementioned challenges and importance of vision sensors-based semantic segmentation in accurate scene understanding and parsing, we accumulate the existing

research contributions and outcomes in this survey. The main research questions that are highlighted in this survey are given as follows. 1) Do the available datasets possess generalization potentials for scene understanding in complex visual scenes? 2) Can the current methods segment complex visual scenes containing uncertainties such as fog and rain, and segment the unstructured information including rough roads and non-smooth pathways for pedestrians? 3) Do the current methods attentively learn from the ongoing scenes and contain events-based scene understanding potentials?

C. Contributions

This survey takes a step ahead in this regard by critically examining the recent state-of-the-art in visual scene understanding using segmentation techniques, with the following main contributions to the ITS community:

1) A thorough introduction to scene understanding, which defines the generic pipeline and explains each of its steps individually. This helps newcomers to the field grasp prior knowledge from all aspects of scene understanding for AD.
2) A discussion and critical analysis of the most relevant papers and datasets arising from the notable research activity on scene understanding witnessed during the last decade.
3) A performance study of current state-of-the-art methods by considering their consumed computational resources and the platforms for which these methods are developed. In the existing literature, some contributions provide their open-source implementations. This review leverages them by executing, analyzing, and comparing the resources consumed by each method. This study allows expanding the target audience of this review towards industrialists with interest in better scene understanding strategies that are functional in real-world environments.
4) A reasoned derivation of future research guidelines based on the analyzed literature, identifying open problems and challenges in this domain, as well as research opportunities that can be explored to address them effectively.

D. Review Methodology

The research articles discussed in our position survey are retrieved using different keywords such as *scene understanding in autonomous vehicles*, *vision-based semantic segmentation in autonomous vehicles*, and *multi-class scene understanding in autonomous driving*. Most of the articles retrieved were purely relevant with some exceptions for multi-modalities methods [31], [32], weak relevance to the investigated topic, for instance, point cloud systems [33], and some outdated articles with relatively old deep learning strategies [34]. Furthermore, the aforementioned keywords are searched in multiple repositories including the Web of Science and Google Scholar to ensure the retrieval of relevant contents. The inclusion criteria ensures that a paper is recognized among the AD experts *i.e.,* the number of citations, where we also analyzed the *Use in the Web of Science* and the classification of citations such as checking whether the concerned paper is cited in most of the articles as a support or in background or general discussion. In Figure 3, the overall distribution is provided, where the statistics indicate that the trending publisher in ITS domain from semantic segmentation understanding perspective is IEEE, followed by non-reviewed pre-prints in ArXiv repositories.

The rest of the manuscript is split into five main sections. Section II highlights the role of segmentation for AD, and explains some featured methods from related literature. Section III explains evaluation metrics in use for segmentation tasks, several loss functions designed for special purposes, and a time complexity analysis of representative methods from the segmentation literature. A list of widely used segmentation datasets are enumerated and described in Section IV, along with an explanatory discussion on the drawbacks and the challenges posed by them. Section V exposes open challenges for scene understanding methods in the AD domain using segmentation modules, and outlines research directions to address them. Finally, in Section VI, we conclude this review with derivations of the whole article and an outlook.

II. Semantic Segmentation for Scene Understanding in AD

The primary objective of semantic segmentation is to annotate each pixel of an input image within a range of predefined classes used while training *i.e.,* defining boundaries of individual entities inside an ongoing scene, assisting in many applications [45]. The dictionary of possible classes varies depending on the dataset and the segmentation task under consideration. Nevertheless, basic objects that are common in most databases used in semantic segmentation literature for AD include humans/pedestrians, different types of vehicles (car, bike, *etc.*), traffic lights, and many more, [46], [47]. Segmenting different types of objects assists the autonomous vehicle decision making. For instance, if a nearby pedestrian is accurately segmented by a deep neural model, it instantly initiates the brake pressing mechanism by considering the distance between vehicle and the pedestrian. This is easily doable using accurate segmentation technique that draws clear boundaries of pedestrian against other objects, contributing to real-time decision making.

Semantic segmentation for scene understanding is mostly performed via RGB cameras. More recently, LiDAR sensors-based methods have achieved significant results in segmenting an outdoor scene for autonomous vehicles [36]. There are major fusion-based techniques that allow RGB data and LiDAR point clouds to interact in a single network for semantic segmentation [40]. However, in this article we specifically focus on RGB sensors-based semantic segmentation methods due to their lower computation cost, high level of applicability, and large field of view. A concise summary about the literature on LiDAR and multi-modalities semantic segmentation is given in Table I. Furthermore, interested

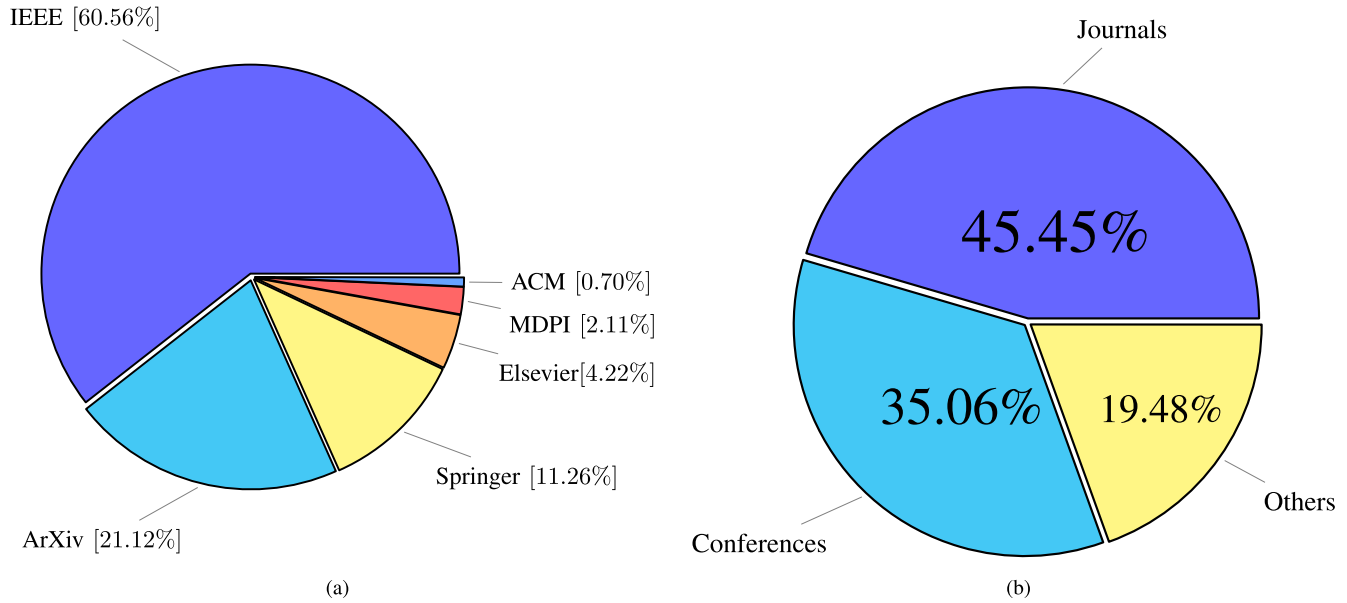


Fig. 3. The overall literature distribution of semantic segmentation driven scene understanding methods for AD. (a) Publisher-wise literature distribution of scene understanding for autonomous driving. (b) Research article type based literature distribution of scene understanding for AD.

readers can refer to a very recent survey on 3D LiDAR data for semantic segmentation available in [48].

We now discuss some prominent segmentation methods featured for AD. Segmentation is widely used in scene parsing, whereas some methods only focus on specific kind of objects such as pedestrian, cars, bicyclist, and lane to incorporate their importance for AD in streets. In order to attribute the desired level of importance to such objects, RAPNet [73] contains importance-aware features selection method to automatically nominate important features for the predicted labels. By contrast, other mainstream methods [60], [65] focus on general objects' segmentation without granting any importance to objects on road or zebra crossing areas. Scene understanding in some methods is performed using segmentation techniques functional in diverse environments with unstructured roads [74], challenging weather [75], outdoor complex conditions [76], and varying illumination [77]. A detailed description of features segmentation methods is given in Table II.

There exists several survey contributions of computer vision research community to cover various major challenges, provide tutorials, and offer future research directions in various subdomains of AD. These surveys are summarized in Table III. As can be observed in this table, scene understanding is not specifically considered to the level of its importance in AD, and there exist very scarce surveys related to scene segmentation. For instance, Xue *et al.* covered scene understanding methods based on events reasoning in their baseline survey [29]. This is the most related survey to our topic, but it is concentrated on events and intention prediction of pedestrians and vehicles rather than on scene parsing and related paradigms. Another recent survey broadly covers road segmentation methods, but without any focus on their concerned challenges with future research directions in the AD domain [78]. To the best of our knowledge, this survey is novel of its kind in the AD literature and is a need of the community working on autonomous vehicles, given the acknowledged importance of scene parsing in this domain.

III. Performance Evaluation of Semantic Segmentation

The performance evaluation of different semantic segmentation models used in AD domain are discussed in this section. Herein, we explain the evaluation metrics, different types of objective functions, analyze the computational complexity, and finally provide quantitative comparisons of deep models. The nomenclature of the used variables is given in Table IV.

A. Evaluation Metrics

Building only a predictive deep segmentation model is not a wise and trustworthy decision for safe AD unless it is tested on unseen data. Most models evaluate their performance on a disjoint set of the same dataset that is used for training, but still the test data are totally new for the trained model. Recently, deep models are being developed with more generalized potentials for unseen data [89]. Deep models for segmentation are evaluated using some common metrics to assess the optimal results against ground truth. Based on the difference between instance and semantic segmentation, different types of evaluation metrics can be used for these tasks, which we review next as follows.

1) Intersection Over Union (IoU): The IoU metric [90], [91] computes the overlapping regions between the predicted model's results $predMask_{input}$ and the ground truth mask GT. It is the simplest metric that essentially counts the number of common pixels using intersection and union as per Equation (1).

$$IoU(predMask_{input}, GT) = \frac{predMask_{input} \cap GT}{predMask_{input} \cup GT}, \quad (1)$$

TABLE I
COMPREHENSIVE DATA TABLE OF EXISTING LITERATURE ON SEMANTIC SEGMENTATION USING LIDAR SENSORY DATA OR FUSION OF RGB AND LIDARS. HEREIN, FEATURED METHODS ARE PROVIDED AND ARE SELECTED BASED ON THEIR ENDORSEMENT AMONG SCENE UNDERSTANDING RESEARCH COMMUNITY FOR AD

Ref.	Main Theme	Functions and Horizon	Remarks
[35]	Stabilization and validation process of the measured position of objects	Multi-Object Tracking, Sensor Fusion, and Motion Compensation	Centered towards multi-target tracking and data association, sensor fusion, white, and black box sensor fusion methods
[36]	Sequential fusion-based 3D pedestrian detection using LiDAR point Cloud and semantic segmentation in automated driving vehicles	Point Painting, Semantic Segmentation Network, Semantic Augmentation, Semantic Features, and Geometric Features	Over-viewed 3D pedestrian detection, semantic segmentation, point cloud augmentation, feature encoding, and fusion schemes
[37]	Inferring semantic information towards an understanding of the surrounding environment for autonomous vehicles	Adaptation models, Laser radar, and Data models	Studied various point cloud semantic segmentation methods for autonomous driving using semantic mapping, domain adaption, semi-synthetic scan simulation, and geodesic correlation alignment
[38]	A deep learning approach for 3D semantic segmentation of LiDAR point clouds	Image segmentation, Three-dimensional displays, Laser radar, Semantics, Cameras, and Feature extraction	A general idea about 2D and 3D semantic segmentation, multi-model 3D semantic segmentation, feature transformation, and fusion
[39]	Focusing on autonomous cars equipped with RGB cameras and LiDAR	Auto-Driving, Robotics, Multi-Sensor Fusion, Perceptual Information, and Spatio-Depth Information	Discussion on a collaborative fusion scheme for autonomous cars equipped with LiDAR and RGB cameras, methods using only RGB camera LiDAR sensor
[40]	Using camera and 3D LiDAR as indispensable devices in modern AD vehicles	AD Vehicles, LiDAR and Camera, LiDAR Segmentation, Contextual Information, and Weak Spatiotemporal Synchronization	An overview of different deep learning approaches using 3D point clouds, LiDAR point cloud semantic segmentation, camera semantic segmentation, and LiDAR and camera fusion semantic segmentation
[41]	Integrating multi-modal perception system for AD	Solid modeling, Laser radar, Three-dimensional displays, and Computational modeling	A debate on approaches using semantic and instance segmentation on RGB images, semantic segmentation on point clouds, spherical projection, and DenseFuseNet framework
[42]	The fusion of LiDAR and camera data for semantic segmentation using a semi-supervised learning technique	Sensors Fusion, Semi-Supervised Learning, and Semantic Segmentation	Centered towards semi-supervised learning, task of AD, fusion techniques, point cloud projection, and sparse semantic masks from 3D bounding boxes
[43]	The autonomous systems can capture and process complementary perceptual information for better detection and classifying objects	Multi-Class 3D Object Detection, LiDAR-Camera Fusion, Multiple Fusion Stages, Point-Wise Fusion, and ROI-Based Feature Pooling	Discussion on LiDAR and multi-sensor based 3D object detection, Point-wise semantic information fusion, Local region fusion, and Multi-label prediction auxiliary regularization.
[44]	An iterative deep fusion architecture for semantic segmentation of 3D point clouds	3D Semantic Segmentation, 2D Semantic Segmentation, Deep Fusion, Convolutional Neural Networks, Point Clouds, and Sensor Fusion	An overview of an iterative deep fusion architecture, semantic segmentation of 3D point clouds, fusion strategies, LiDAR, and camera segmentation

where $predMask_{input}$ is the mask of labels predicted for each pixel of the input image, and GT is the ground truth mask that should be predicted by an ideal segmentation model. In case of multiple classes (as it often occurs in the related literature), the IoU score is computed for each class individually followed by its global average over all classes, giving rise to the so-called mean IoU. As this method is based on Jaccard and Dice coefficients, it is also referred to as Jaccard Index.

Computing IoU over the output of instance segmentation models is complicated, as it produces multiple masks for each object inside an input image. Therefore, it becomes similar to object detection evaluation with the only difference being the bounding boxes comparison in the object detection problem, which is replaced by the masks comparison in instance segmentation.

2) Pixel Accuracy for Semantic Segmentation: Another commonly used metric is the pixel accuracy $PixelAcc$ [57], which reports the percentage of correctly classified pixels in an input image when correspondingly compared to the ground truth mask, as formulated in Equation (2).

$$PixelAcc(\ell) = \frac{TP(\ell) + TN(\ell)}{TP(\ell) + TN(\ell) + FP(\ell) + FN(\ell)}, \quad (2)$$

where $TP(\ell)$, $TN(\ell)$, $FP(\ell)$, and $FN(\ell)$ respectively denote the number of true positives, true negatives, false positives, and false negatives measured over the image, assuming that pixels of label ℓ are given value 1 and 0 otherwise. As in IoU, it is also computed individually for every class, and globally for all classes of a given dataset. For a single-class representation with comparatively smaller coverage in an image, this metric is biased as it only reports on the identification of pixels in an image where a class (positive class) is not present.

B. Special Loss Functions for Semantic Segmentation

In general, various factors may affect the learning potentials of a certain Machine Learning model. The loss function is among the most important ones in neural computation, as it quantitatively evaluates the model's predictions during training and improves the performance via gradient updates and back

TABLE II

COMPREHENSIVE ANALYSIS OF EXISTING LITERATURE ON SCENE UNDERSTANDING METHODS FUNCTIONAL FOR AUTONOMOUS DRIVING. HEREIN, FEATURED METHODS ARE PROVIDED AND ARE SELECTED BASED ON THEIR ENDORSEMENT AMONG SCENE UNDERSTANDING RESEARCH COMMUNITY FOR AD

Ref.	Method	Dataset/Availability	Implementation	Remarks	Traffic flow/Scenario		
					High	Moderate	Speed
[49]	Encoder with parallel dilated convolutions and decoder using transposed convolutions	Cityscapes [50]	N/A	Yes/	✓		Controlled
[51]	Two stages *i.e.,* semantic labelling and points re-projection for scene understanding	Semantickitti [47], KITTI [52], and [46]	N/A	Yes/Semantically segment the scene and combine with LiDAR point data to obtain (segment) drivable space	✓		Controlled
[53]	A benchmark dataset	Data from autonomous vehicle sensors suite	–	Yes/A very large scale data with 7x as many annotation and 100x as many images as the pioneering KITTI datasets [52], [46]	✓		Controlled
[54]	Deep reinforcement learning algorithm that takes transitions of all vehicles in sensors' range to have a global reward function	HidghD dataset [55]	N/A	Indirectly related/Focused on driver's change in lane detection behavior		✓	High (highways/ countryside area)
[56]	A pedestrian location perception model for complex driving scenes	CityScapes dataset [50]	https://github.com/espci/p_spn/blob/master	Indirectly related/ Focused on pedestrian location information for scene understanding	✓		Controlled
[57]	Geometry (depth) and motion (optical flow) with semantics learning	KITTI and CityScapes [50] datasets [52]	https://github.com/CVLAB-Unibo/omeganet	Yes/The system is analyzed using GPU and lower-power embedded system.		✓	Variable (urban roads/ highways)
[58]	Simple Unet inspired architecture for occupancy grid computation using radar data	Nuscenes [53]	https://github.com/liat-s/radar_occupancy_grid/ (code not uploaded yet)	No/Different type of data, not RGB camera data	✓		Controlled
[59]	Three levels of outputs: semantic, instance labels, and depth image	CityScapes [50]	N/A	Yes/Three type of losses are combined.	✓		Controlled
[60]	Future semantics, geometry, and motion inside scene to control autonomous vehicle	CityScapes [50], Mapillary Vistas [61], ApolloScape [62], and Berkely Deep Drive [63]	No/Project page: https://wayve.ai/blog/predicting-the-future/	Yes/Their method predicts future semantics.		✓	Variable (urban roads/ highways)
[64]	Focused on detecting free drivable segmented area, objects' distance from camera and orientation	Own dataset/No	N/A	Yes/Identifies closest obstacle in each direction and free drivable area delimitation	✓		Controlled
[65]	Scene understanding using joint object detection and semantic segmentation	COCO [66], VOC 2007 [67], and VOC 2012 [68] /Yes	https://github.com/dvornikita/blitznet	Yes/Identifies closest obstacle in each direction and free drivable area delimitation	✓		Controlled
[69]	Honda Research Institute Driving dataset and a novel driver behavior understanding annotation method	N/A /Yes	https://usa.honda-ri.com/HDD	No/Driver behavior reasoning	✓		Controlled
[70]	Formed DNN based clustering tags of image regions with common appearance and using tags for optimal segmentation	CityScapes [50] and CamVid [71]/Yes	N/A	Yes/Efficient semantic segmentation for AD	✓		Controlled
[72]	A multi-label model for diverse scenes recognition	New dataset, [68]/Yes	N/A	Yes/Efficient semantic segmentation for AD		✓	Variable (urban roads/ highways)
[73]	Streets scene understanding model that focuses on common objects using importance-aware feature selection mechanism	CityScapes [50] and CamVid [71]/Yes	N/A	Yes/Segmentation focused on street scenes aware objects		✓	Variable (urban roads/ highways)

propagation until the the specified number of epochs. There are multiple loss functions for segmentation tasks. Furthermore, some research works have hitherto proposed to improve the segmentation performance further by defining modified/hybrid versions of these loss functions. Common loss functions can be found in [92], whereas advanced type of loss functions are given below with their respective mathematical definitions:

1) Weighted Binary Cross Entropy: It is a variant of cross entropy loss function that is widely used in many computer vision problems. It is defined as the difference measure of two probability distributions (y and $\widehat{y}$) of corresponding inputs [93], [94]. In this case, β is used for balancing among false positives and negatives.

$$wBCE(y, \widehat{y}) = -\beta \cdot y \log(\widehat{y}) + (1 - y) \log(1 - \widehat{y}), \quad (3)$$

where $\widehat{y}$ is the output of the segmentation network for a given pixel, and y is its ground truth, and the images and labels weights are computed using zeros and ones.

2) Balanced Cross Entropy: In this alternative formulation of the loss function [95], [96], positive and negative samples are weighted as follows:

$$wbBCE(y, \widehat{y}) = -\beta \cdot y \log(\widehat{y}) + (1 - \beta) \cdot (1 - y) \log(1 - \widehat{y}), \quad (4)$$

hence inserting a complementary weight for negative samples.

3) Focal Loss: Focal loss FL is a well-established loss function that can be used in case of imbalanced data [97], which also occurs frequently in segmentation problems.

TABLE III

DETAILED DESCRIPTIONS OF SOME REPRESENTATIVE SURVEYS IN THE ITS LITERATURE, SORTED IN TERMS OF THEIR RELEVANCY TO THE PRESENT SURVEY

Ref.	Year	Main theme	Domain	Remarks
[79]	2015	Focusing on motion planning techniques and navigation	Graph search-based planners, sampling-based planners, interpolating curve planners, and numerical optimization approaches	Studied various motion planning methods for self-driving vehicles, highlighted motion planning, and path selection methods for autonomous urban vehicles
[80]	2016	Motion planning and control techniques for autonomous urban vehicles	Route planning, behavioral decision making, motion planning, and vehicle control	Over-viewed only traditional approaches for motion planning and briefly discussed the major challenges with recommendations in future work section
[81]	2017	Simultaneous localization and mapping survey, focusing on current trends for autonomous vehicles	Evaluation of SLAM technology in autonomous driving, single-vehicle SLAM, multi-vehicle slam, centralized SLAM, and decentralized SLAM	Discussed conventional hand-crafted methods and few deep learning techniques, and Deep learning-based methods are recommended over handcrafted methods
[29]	2018	Events reasoning based scene understanding survey	Scene representation, events detection, intention prediction for AD	Centered towards events reasoning (pedestrian and vehicle events), traffic saliency detection, and intention prediction (long- and short-term). Events and intentions' prediction evaluation discussion is provided. Not specific towards the recent advancement of deep models for scene understanding. Comparatively old literature is studied in this survey.
[82]	2018	Coverage of intra- and inter-vehicle networking and communication technologies in AD	Vehicle identification, emergency warning, distance control, road hazard warning, and cooperative precise localization	Highlighted the importance of wired and wireless communication technologies for reliable autonomous driving, and new trends of networking technologies in autonomous vehicles
[83]	2018	Investigation of state-of-the-art vehicle localization techniques for AD	GPS/IMU-, camera-, radar-, LiDAR-, ultrasonic-based vehicle localization	Over-viewed conventional sensors- and vision-based vehicle localization and various challenges are highlighted with recommended solution
[84]	2019	Interaction between autonomous vehicles and pedestrians, recognizing the pedestrian's intention, and behavior	Pedestrian detection and tracking, pedestrians intention estimation, pedestrians decision and behavior prediction	Discussed the importance of practical approaches for communicating with pedestrians, understanding the pedestrian's behavior, intention, and decision based on traffic characteristics
[85]	2019	The applications of radar signals for road safety in autonomous driving	Range estimation, velocity estimation, angle estimation between two cars or between car and pedestrian towards safe AD	Presented state-of-the-art techniques, phase noise estimation methods, and investigated the performance of radar in environmental phase noise.
[86]	2019	Efficiency of radar signals for the detection of far-small and close-large objects in autonomous vehicles	Extended target detection, multi-path mitigation, angular super-resolution, clustering, and multiple target tracking	Discussed the traditional signal processing approaches for AD and future research directions of signal processing for practical roadway scenarios
[87]	2019	3D object detection methods for AD applications	Predictions of 3D object bounding boxes, 2D bounding boxes prediction on the image plane, and projection of point clouds into a 2D images	Over-viewed the object detection methods, the commonly used sensors, datasets, the type of depth data user for 3D road objects detection, and detailed comparative analysis of 2D and 3D object detection for autonomous driving
[78]	2020	Investigated the current challenges and future directions for safe AD	Road segmentation, pedestrian detection, lane segmentation, drowsiness detection, collision avoidance, and traffic sign detection for driving safety in autonomous vehicles	Covered only the most recent deep learning approaches for safe autonomous driving, detailed analysis, experimental evaluation, key findings, major challenges, and their recommended solutions
[88]	2020	A comprehensive survey of computer vision based autonomous vehicles	Benchmark AD datasets, object detection, object tracking, semantic segmentation, and semantic instance segmentation	Presented numerous computer vision-based solutions for AD, highlighted the current challenges in various domains of computer vision-based approaches including motion estimation, object tracking, and end-to-end object segmentation
Ours	2022	A study of deep learning models based scene understanding/segmentation methods	Scene segmentation models, datasets, evaluation, and loss functions for AD	We introduce scene segmentation, their open challenges, discuss its role for AD, and explain future directions with respect to scene understanding.

Following the previous notation, the focal loss is given by:

$$FL(pt) = -\alpha_t \cdot (1 - p_t)^{\gamma} \log(p_t), \tag{5}$$

where p_t is the probability that the model predicts for the ground truth object, $\gamma > 0$ is an parameter that permits to grant more or less relative weight to misclassified examples, and $\alpha_t \in [0, 1]$ is set to account for the presence of class

TABLE IV
LIST OF USED VARIABLES

Variable	Description
IoU	Intersection over union
$predMask_{input}$	Predicted mask label per pixel of an input image
GT	Ground truth segmentation mask
PixelAcc	Pixel accuracy metric, estimating pixel level accuracy for semantic segmentation
ℓ	l_{th} pixel of segmentation mask, 0 or 1 value
TP	True positive
TN	True negative
FP	False positive
FN	False negative
BCE	Binary cross entropy
β	Balancing factor for unbiased trade-off between false positive and false negative
y	The ground truth segmented pixel
$\widehat{y}$	The predicted segmented pixel
w	Weight value (having values 0 or 1)
FL	Focal loss
P_t	Probability of segmentation model for GT
α_t	Parameter with 0 or 1 output, predicts class imbalance
Mean IoU	Averaged intersection over union
mAP	Mean average precision

imbalance or instead, tuned as another hyper-parameter of the overall model.

4) Others: There are many other types of loss functions[1] used in specific cases for segmentation problems, such as region-based losses [98]. Among them we underscore the prevalence of studies using the Dice loss, which gets inspired by the Sørensen–Dice coefficient (namely, a measure of similarity between images); the Tversky loss, which extends the Dice loss with a β coefficient to weight differently false negatives and positives; the shape-aware loss for better addressing the segmentation of challenging objects; the Hausdorff distance loss [99], [100]; and the combo loss, which blends together the binary cross-entropy loss for curves smoothing effect and the Dice loss for class balancing problems. We again refer to [92] for a detailed mathematical compendium of these loss functions.

C. Time Complexity Analysis

In order to illustrate the current performance levels of segmentation models used for AD, we now report the results of some featured deep segmentation models. The overall report of running time of these models is given in Table V. Some of the model's time complexity indicators are reported from their methods, whereas in other cases we have run the reported models from their publicly available repositories using our experimental resources. The system's configuration for CPU includes an Intel(R) Core i7-7700 CPU@3.60 GHz processor running on Windows 10 operating system, while the GPU used in experimentation is a NVIDIA GeForce GTX 1060 with 6 GB graphics memory. Table V also shows the predictive performance of the models (when available) over three different datasets, as well as the size of the trained models (measured in MB).

[1]https://cnvrg.io/semantic-segmentation/ (accessed on April 21st, 2021).

TABLE V
PERFORMANCE AND TIME COMPLEXITY ANALYSIS OF FEATURED SEGMENTATION MODELS. AN UPWARD ARROW DENOTES THAT THE HIGHER THE VALUE, THE BETTER THE MODEL (AND VICE VERSA FOR THE DOWNWARD ARROW)

Model	Mean IoU (↑)			Size (MB) (↓)	FPS (↑)		Time (ms) (↓)
	Pascal	CityScapes	KITTI		CPU	GPU	
[57]	-	82.95	75.84	118	3	57.4	-
[10]	79.70	70.42	-	439	2.5	30	-
[107]	-	80.35	56.43	165	-	-	166
[65]	75.72	-	-	705	-	24	200
[108]	64.86	-	-	213	-	24	-
[109]	-	50.73	-	171	-	-	10.43
[102]	-	71.31	-	11.9	-	81.9	-
[110]		69.14	-	705	-	24	-
[111]	-	72.0	-	8.5	-	51.7	-
[112]	-	75.23	-	2.43	-	-	136

TABLE VI
QUANTITATIVE ANALYSIS

Dataset	Ref.	IoU/mIoU	Pixel Accuracy	mAP	FPS	Settings
CityScapes [50]	[9]	79.4	-	-	3.5	Nvidia Titan X
	[26]	80.4	-	-	46.8	Nvidia Titan X
	[27]	74.4	-	-	65.5	Nvidia Titan Xp
	[28]	67.70	-	-	30.3	Maxwell Titan X
	[49]	59.8	-	-	16.7	Nvidia Tegra X1
	[56]	0.68	-	-	22	Nvidia GTX1080Ti
	[57]	82.92	92.50	-	15	NVIDIA Titan Xp
	[59]	80.40	-	-	21	Pascal Titanx
	[60]	0.464	-	-	-	-
	[70]	53.70	-	-	-	-
	[73]	92.30	-	-	8.5	NVIDIA Titan X
	[102]	70.3	-	-	100	NVIDIA Titan X
	[107]	82.10	-	83.80	-	-
	[110]	67.40	-	-	45.1	Single 1080Ti
	[111]	70.5	-	-	51.7	NVIDIA GTX 1080Ti
CamVid17 [71]	[27]	68.70	-	-	65.5	Maxwell Titan X
	[28]	67.10	-	-	27.8	Maxwell Titan X
	[70]	30.40	-	-	-	-
	[102]	64.70	-	-	120	NVIDIA Titan X
	[110]	65.0	-	-	38.7	Single 1080Ti
COCO [66]	[27]	31.30	65.50	-	65.5	NVIDIA Titan Xp
	[28]	29.10	-	-	35.7	Maxwell Titan X
	[65]	53.50	-	34.10	-	-
S.KITTI [47]	[51]	52.20	-	-	92	NVIDIA Drive AGX
VOC07,12 [67]	[65]	75.70	-	83.60	19.5	Maxwell Titan X
	[107]	64.86	91.20	-	-	-

The world's leading AV chips including Intel Ponte Vecchio, NVIDIA A100, Tesla D1, Huawei Ascend 910, and Google TPU (v1, v2, v3), have achieved mass production for applications such as 2D/3D fusion annotation and semantic segmentation training [101]; however, the time complexity of the analyzed methods running over CPU indicates that the current neural architectures still need to focus on lowering the time complexity and energy consumption. The highest frames per second (FPS) among these methods is achieved by [57], that is 3 frames per second for CPU. When deployed over a GPU, the best FPS score is 81.9 frames per second achieved by [102]. In real-world environments [103], devices are severely resource-restricted [104], such as Raspberry-pi, Jetson Nano, and Google Board. Executing such huge models over these devices is a challenging task. Therefore, much attention is required in terms of time complexity towards enabling the execution of these models over energy-limited devices functional in Internet of Things setups [105], [106].

D. Quantitative Analysis of Scene Segmentation Methods for AD

This section elaborates on the quantitative empirical analysis of road scene segmentation methods surveyed in this paper

TABLE VII

STATISTICAL OVERVIEW OF THE SCENE UNDERSTANDING DATASETS USED FOR AUTONOMOUS DRIVING. * REPRESENTS A DATASET WITH ADVERSE WEATHER CONDITIONS

Dataset	Year	No. of Scans	Resolution	Data format		Classes	Horizon
				2D-Object	3D-Scan		
VOC (2007 and 2012)	2010	22,531	-	✓		20	Detection
KITTI	2012	14,999	1240×376		✓	28	Segmentation
COCO	2014	330,000	640×480	✓		80	Detection
CityScapes	2016	5000	1024×2048	✓		30	Segmentation
Mapillary Vistas	2017	25,000	1920×1080	✓		66	Segmentation
HighD	2018	1,530,000	4096×2160	✓		-	Detection
ApolloScape	2019	143,906	3384×2710		✓	-	Segmentation
SemanticKITTI	2019	43,552	-		✓	28	Detection
NuScenes	2020	4000	1600×900		✓	23	Detection
Barkely Deep Drive	2020	10,000	-	✓		-	Segmentation
ACDC*	2021	4006	-	✓		19	Segmentation

and that empower the automation of AD. For the quantitative assessment, we inspect the performance of every method for road scene segmentation using three evaluation metrics: Mean Intersection over Union (mIoU), Pixel Accuracy, and Mean Average Precision (mAP). Furthermore, computational efficiency is accounted for by reporting the FPS achievable by each method in inference time. The detailed quantitative results in terms of the aforementioned metrics are given in Table VI. Results across different scene segmentation datasets (including Cityscapes, CamVid17, COCO, SemanticKITTI, and VOC) are reported from the literature and compared based on their results.

From the reported results in Table VI, it can be noticed that, among all methods evaluated over the Cityscapes dataset, the approaches proposed in [26], [27], and [102] attain a balanced trade-off between accuracy (in terms of mIoU, Pixel Accuracy, and mAP) and efficiency for real-time applications (in terms of FPS). By contrast the reported results over the CamVid17 dataset evince a better segmentation performance of the methods contributed in [27] and [102]. Among these three focused methods, [27] scores best in terms of mIoU and mAP values, with superior FPS, which are 31.30, 65.50, and 65.5, respectively. The reported results over the SemanticKITTI dataset indicate a better performance of the method in [51], achieving well-balanced mIoU and FPS scores, i.e., 52.20 and 92, respectively. Finally, the method in [65] performs comparatively better than the one in [107], by offering best values of the mIoU, mAP, and FPS scores (75.70, 83.60, and 19.5, respectively).

IV. DATASETS

Many datasets are nowadays available for segmentation tasks, where some of them are related to semantic segmentation and others are introduced for instance segmentation. Representative datasets in the segmentation literature particularly those designed for AD are discussed in detail in the subsequent sections and their detailed statistics are given Table VII.

A. KITTI

KITTI [46] is a 3D vision benchmark data containing outdoor stereo images of road scenery along with its corresponding 3D laser scans. The 3D image data is acquired by two high resolution stereo cameras (gray scale and color), advanced OXTS RT 3003 localization system that combines global positiong system (GPS), global navigation satellite system (GLONASS), inertial measurement unit (IMU), and real time kinematic (RTK) correction signals. It also contains Velodyne HDL-64E laser scanner, mounted on the top vehicle to produce 3D points for the captured scenes in real time. The deployed stereo cameras are first calibrated and then synchronized with a localization system and a laser scanner to generate accurate ground truth data.

The dataset comprises a total of 14999 RGB stereo image pairs (including both image and its corresponding ground truth), with a resolution of 1240 × 376 pixels. The entire dataset is partitioned into a training (7841 samples) and a test set (7518 samples). The training set is further split into two subsets, namely, train (3712 samples) and test set (3769 samples), and the latter is used mainly for validation purposes.

B. SemanticKITTI

SemanticKITTI [47] is a large-scale outdoor scene dataset constructed for point cloud semantic and panoptic segmentation of road scenery, including residential area, city traffic, and highways. It comprises a total of 43552 point-wise re-annotated 3D scans generated with automotive LiDAR sensor for the KITTI Vision Odometry Benchmark dataset [46]. This dataset has a total of 22 distinct sequences split into training-validation and test subsets. The training-validation set consists of 23,201 3D scans from sequences 0 to 10, while the test set comprises of 20,351 3D scans from sequences 11 to 21.

Unlike Paris-Lille-3D [113] and Wachtberg [114] datasets, which only contain the aggregated 3D scans of the complete sequence captured with the same type of sensors, SemanticKITTI provides the individual point cloud of the entire captured sequence of road scenery. Thus, it enables the performance evaluation of semantic segmentation based on multiple consecutive scans.

C. HighD

The HighD dataset [55] contains around 110,000 refined trajectories of different vehicles, including cars and trucks. Those trajectories are captured from drone videos recorded at a resolution of 4096 × 2160 pixels and 25 FPS over German highways. For each particular vehicle trajectory, the dataset provides trajectory ID, speed, acceleration, longitudinal coordinate, distance to the leader, and ID of the current leader. These trajectories are widely used to analyze the driving behavior of car-following drivers using computer vision algorithms. The dataset includes 60 videos of 17 minutes on average captured in 6 different locations, depicting a road portion of around 420 meters in length. All videos are captured in sunny and clear weather conditions, from 8 AM to 5 PM, thereby minimizing the efforts required for video stabilization and other post-processing operations.

The dataset includes four different files for each captured video, including three CSV files and the visual aerial view of the highway. The first file contains the information about traffic signs, driving lanes, speed limit on each specific lane,

and location of the site. The vehicle class, vehicle dimensions, mean speed, and driving direction is given in the second file. The third file provides the detailed information such as speeds, lane position, accelerations, and description of adjacent vehicles per frame.

D. CityScapes

CityScapes [50] is a high-quality pixel-level semantic segmentation dataset for urban street scene understanding, collected in around 50 cities in Germany and neighboring countries. The dataset provides 5,000 pixel-level annotated images of resolution 1024×2048, depicting complex urban scenes captured in different weather conditions, varying background, and scene layout. As compared to other benchmark datasets for street scene understanding [46], [47], [55], the CityScapes dataset surpasses the previous efforts in terms of variety, size, scene complexity, and annotation richness.

To discriminate the semantic representation of each particular object in the captured image, data is annotated with 30 different categories. For semantic segmentation task, the entire dataset is split into four separate subsets including 2,993 training images, 503 validation images, 1,531 test images, and 20,021 auxiliary images. The training, validation, and test image sets have high-level refined annotations, while the auxiliary set of images contains coarse annotations.

E. Nuscenes

NuScenes [53] is a large-scale 3D object detection dataset recently introduced for driving scene understanding in AD. The dataset is collected in Boston (South Boston and Seaport) and Singapore (Holland Village, Queenstown, and One North) using moving car equipped with a suite of specially designed sensors. The car-mounted suite includes 13 sensors: 6 RGB cameras with 1600×900 resolution and 12Hz capture frequency, 5 long-range radar sensors operating at 77 GHZ with 13Hz capture frequency, 1 LiDAR sensor with 20Hz capture frequency, and an IMU sensor. All sensors are precisely synchronized with each other to obtain high-quality data and better cross-modality between visual and sequential data.

The dataset consists of 1000 driving sequences, where each sequence is 20 seconds long. Data are annotated by experts into 23 object classes (*i.e.,* Car, Truck, Human, and Bicycle *etc.*), where each object class is further categorized into 10 different sequence classes based on the semantic differences between the sequences. For training and inference, the dataset is divided into 700, 150, and 150 annotated sequences for training, validation, and testing, respectively. Each sequence comprises 40 frames, offering a 360^o view of the surrounding scenery.

F. Mapillary Vistas

The Mapillary Vistas [62] is one of the largest and challenging street-level scene segmentation datasets for pedestrian and traffic-related scene analysis. The dataset contains 25,000 high quality (8.6 Pixels) outdoor scene images of resolution 1920×1080 captured from all over the world at different conditions concerning lightning, season, weather, and daytime. Images are captured by the sidewalk pedestrians as well as from the moving cars with various image acquisition devices including smart phone cameras, action cameras, tablets, and professional cameras. To prepare the data for supervised learning-based scene segmentation, data are annotated into 66 distinct object categories with additional 37 classes with instance-specific labels.

The Mapillary Vistas dataset is 5 times larger than the benchmark CityScapes dataset [50], providing fine-grained annotated data generated by 69 expert annotators with polygon style for delineating each specific object in the image. For semantic segmentation learning task, the dataset is split into three subsets of images namely training, validation, and testing, having a total of 18,000, 2,000, and 5,000 annotated images, respectively.

G. ApolloScape

ApolloScape [115] is an extensive street-level road scene dataset recently released for a variety of self-driving applications including car instance segmentation, 3D map construction self-location, scene parsing, lane segmentation, scene trajectories, and detection-tracking. The dataset contains 143,906 frames of resolution 3384×2710 pixels, with good quality ground-truth data, comprising pixel-level semantic segmentation, pose information, and 3D point clouds of captured scene. Compared to the existing publicly available datasets (*i.e.,* KITTI [46] or the Mapillary Vistas [62]), ApolloScape comprises almost 15 times more data with rich labeling in terms of holistic semantic dense point for each scene.

The images and depth data in the dataset are acquired with car-mounted sensors deployed over various cities of China under different weather (cloudy and sunny), lightning (day, night, noon), and traffic conditions (rush and non-rush hours traffic with pair of stereo images). The suite of car-mounted sensors includes one VMX-CS6 camera system with two front cameras having a resolution of 3384×2710 pixels, two VUX-1HA laser scanners with range of 1.2m to 420m and 360^o FOV, a measuring head device with IMU/GNSS (heading accuracy 0.015^o, position accuracy $20\sim$ 50%, and roll and pitch accuracy 0.005^o). During data recording, the vehicle drives with a speed of 30 km per hour, whereas the mounted cameras are triggered every 1 meter.

H. Berkely Deep Drive

The Berkely Deep Drive dataset [116] is a large-scale dataset composed by diverse driving videos and GPS/IMU data for road scene understanding including drive-able area segmentation, road objects detection, instance segmentation, and lane mark detection. The dataset includes around 10,000 hours of driving stream depicting visuals of towns, highways, and rural areas of San Francisco Bay Area, New York, and other cities of USA in varying weather and lightning conditions. Besides the video data, the dataset also provides GPS/IMU driving trajectories for location tracking, recorded with GPS, IMU, gyroscope, and magnetometer sensors. The dataset provides image-level annotations for a variety of driving scene understanding tasks. Object detection annotations include traffic light, traffic sign, bus, person, motor, bike, truck, car, train,

and rider. The instance segmentation annotations contain car, road, pedestrian, person, footpath, and traffic boards etc.

I. COCO

The Common Objects in Context (COCO) dataset [66] is one of the predominant databases released by Microsoft, widely used for object detection, semantic and object instance segmentation, and object captioning. The dataset embeds 330,000 images with more than 200,000 labeled instances, 250,000 persons with key points, human pose estimation, and 1,500,000 object instances categorized in 80 distinct classes. The image data is collected from different sources including relevant object images from the PASCAL VOC dataset [67] and the Flickr site uploaded by amateur photographers with search-able keywords. The entire dataset is collected and annotated for object detection, instance segmentation, and image captioning using an interface specifically designed for hired expert annotators.

Originally the COCO dataset is released into two parts: the first part of the dataset was released in 2014, where the second part of the dataset was introduced in 2015. The first part comprises three subsets of images including 82,783 training, 40,504 validation, and 40,775 testing images. Likewise, the second release of the dataset comprises 165,482, 81,208, and 81,434 images for training, validation, and testing, respectively.

J. VOC (2007 and 2012)

The PASCAL VOC (Visual Object Classes) [67] is one of the most challenging datasets publicly available and is used for image classification, object detection, and image segmentation. Similar to the COCO [66] dataset, the VOC dataset is released into two parts: VOC 2007 and VOC 2012. The VOC 2007 release contains a total of 9,962 images and their corresponding annotations split into three subsets: 2501 training, 2510 validation, and 4951 testing images. The VOC 2012 release includes 22,531 images divided into three subsets of 5,717, 5,823, and 10,991 images for training, validation, and testing, respectively. The dataset is captured from two different sources (flickr photo-sharing website and the Microsoft Research Cambridge database).

All images of the VOC 2007 and VOC 2012 datasets are annotated with two distinct attributes, i.e., object class and bounding box, which denote the object type and the coordinates of the object location. Both datasets contain 20 classes, where each class contains a varying number of images. However, each class contains at least 500 images, depicting common objects such as cat, dog, person, car, and bike. For each of these categories, a comprehensive set of images is supplied, each having semantic richness and significant variability concerning to object size, illumination, pose, occlusion, orientation, and position.

V. SCENE UNDERSTANDING IN AD: CHALLENGES AND DIRECTIONS

The datasets introduced above possess a wide variety of objects, with some of them posing least importance towards decision making of an autonomous vehicle such as sky and buildings. Mainstream research contributed nowadays is centered towards favorable daytime scenes for semantic segmentation, with sufficient illumination and supportive weather conditions. Many car companies and Original Equipment Manufacturer in industry have access to a high volume of data; however, they are not keen to share their data publicly, mainly due to IP, industrial competitions, and General Data Protection Regulations (GDPR) concerns. Consequently, lack of sufficient labelled data for accurate scene understanding in dynamic weather conditions with varied illumination conditions, such as night time [117], smoggy situations, and edge cases remains a challenging task for AD research.

This research niche is among the challenges that are still insufficiently addressed by the community to date. In this section we offer our critical views on the current status of scene understanding in AD, summarizing them in a set of challenges together with a prescription of the research directions that can help the community step further and overcome them effectively.

A. Open Challenges

Although significant research has been done and AD industry is widely growing but still there are several open challenges to achieve perfectly intelligent AD, demanding researchers' attention. These challenges are discussed individually with supported references from the related literature.

1) Salient Objects Consideration: While much work has been done in the field of segmentation, very less attention has been paid to objects' distinction based on safety levels or priorities. For instance, a segmentation model only segments humans in an ongoing scene without any consideration of their location or their movement speed, which can be useful to control the autonomous vehicle and avoid accidents. There are various challenges while considering an object's location during segmentation. For instance, the distance of the object from the autonomous vehicle, where the closest distance can be segmented as the highest risky level and the vehicle needs to take actions accordingly. Similarly, an object using zebra crossing and another one walking on roadside can be prioritized differently [73]. Furthermore, motion of the objects [118] from or towards the autonomous vehicle is also an open issue to be faced by future deep learning models for scene segmentation. Object's motion towards the autonomous vehicle with higher speed segmentation map needs quicker actions and vice versa.

2) Coarse-Structured Information: Most of the datasets introduced in AD literature for segmentation are recorded in normal and well-structured infrastructures of advanced cities. The currently developed deep learning models may achieve best results[2] over structured datasets [50], but generalize poorly in many unstructured environments, as given in a sample scenario in Figure 4. For instance, an online challenge NCVPRIPG-2019 focused on unstructured road data recorded

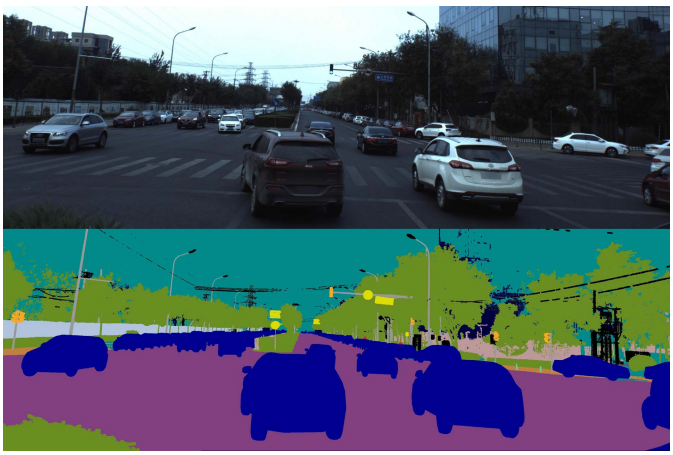
[2]https://www.cityscapes-dataset.com/benchmarks (access: April 8th, 2021).

in India.[3] The highest mean IoU achieved so far in this competition is 0.6276 over the testing set, which reflects the enormous difficulty of achieving models with good generalization properties in complex scenes. This aspect of AD demands further attention in terms of data collection, as well as the inclusion of new and effective representation mechanisms in deep learning models.

3) Uncertainty-Aware Decisions: A largely overseen aspect of scene understanding and AD decision making thereof is the confidence under which models elicit their predictions over the input data. The fact that the vehicular surroundings are inherently uncertain ecosystems seem not to have persuaded the community to delve into this matter, stepping aside current methodological trends centered exclusively on predictive scores. Fortunately, confidence estimation has grasped the attention of the community recently (see *e.g.* [119], [120], [121] and references therein included). Nevertheless, elements from evidential deep learning [122], Bayesian formulations of deep neural networks [123], simpler mechanisms to approximate the output confidence of neural networks (e.g. Monte Carlo dropout [124] or ensembles [125]) and other assorted methods for uncertainty quantification [126] should be progressively incorporated as an additional yet crucial criterion for decision making. This is specially important when dealing with complex environments, in which the lack of data that can fully represent any possible scene induces a large amount of epistemic uncertainty in the output of the model. Without confidence being considered as an additional factor for AD, or with current studies focused solely on predictive and/or computational efficiency aspects, there will be no guarantees that new scene segmentation models upsurging in the scientific community are of practical use and can be transferred to industry.

(a) A road without any specific boundaries or lanes.

(b) An unstructured road with groceries and food items over the main lane. Pedestrians seem walking without any zebra cross or traffic lights.



(c) Sample segmentation ground truth image of a well structured road.

Fig. 4. Sample images with segmentation ground truth from coarse-structured road of India Driving Lite Dataset [74] and CVPR-2018 autonomous driving challenge dataset[4]. The images in (a) and (b) unleash several challenges for a deep learning segmentation model, whereas the image in (c) seems to be easily segmented for scene understanding.

B. Future Directions

The aforementioned challenges and our literature analysis suggest a number of research opportunities for advancing over the current state-of-the-art in vision-based semantic segmentation-assisted scene understanding for AD. We herein offer our envisioned directions:

1) Explainable AD: Deep segmentation models emerging from the AD literature generate their output without eliciting any explanations of how it applied an action during the drive, associating the model's decision with certain complications. If a certain non-explanatory decision of an autonomous vehicle led to an erroneous behavior, causing accidents and traffic irregularities would be problematic from the legal perspective. Explanations of the model's decision are necessary for AI-based decisions to be verified, interpretable, and accountable.

Recently, many deep models are there to explain the generated output [127], that could be applied in AD domain to explain the contributions of a model in a driving decision. Considering achievements so far in explainable Artificial Intelligence (XAI, [128]), AD can harness the myriad of post-hoc XAI techniques available for generating explanations. However, such produced explanations may not suffice in practice as their limited scope may not demonstrate the overall interpretation of a model, but rather provide a correspondence between what the model observes in an input to predict their output. Further research is extensively needed in this direction to produce an enriched narrative connecting vehicular perception to automated actions, as we elevate gradually towards realizing the highest level of AD.

2) Towards Video Segmentation for AD: Semantic segmentation using frame-based visual data has achieved considerable attention, with major improvements in the last two years. Although there are significantly robust techniques for frame-level segmentation, they are still mostly designed for achieving better accuracy levels, compromising their computation efficiency. Therefore, when image-based segmentation is employed in AD, it results in large processing latencies

[3]https://cvit.iiit.ac.in/ncvpripg19/idd-challenge/ (access: April 8th, 2021).

[4]CVPR 2018 WAD Video Segmentation Challenge, https://www.kaggle.com/c/cvpr-2018-autonomous-driving (accessed: April 15th, 2021).

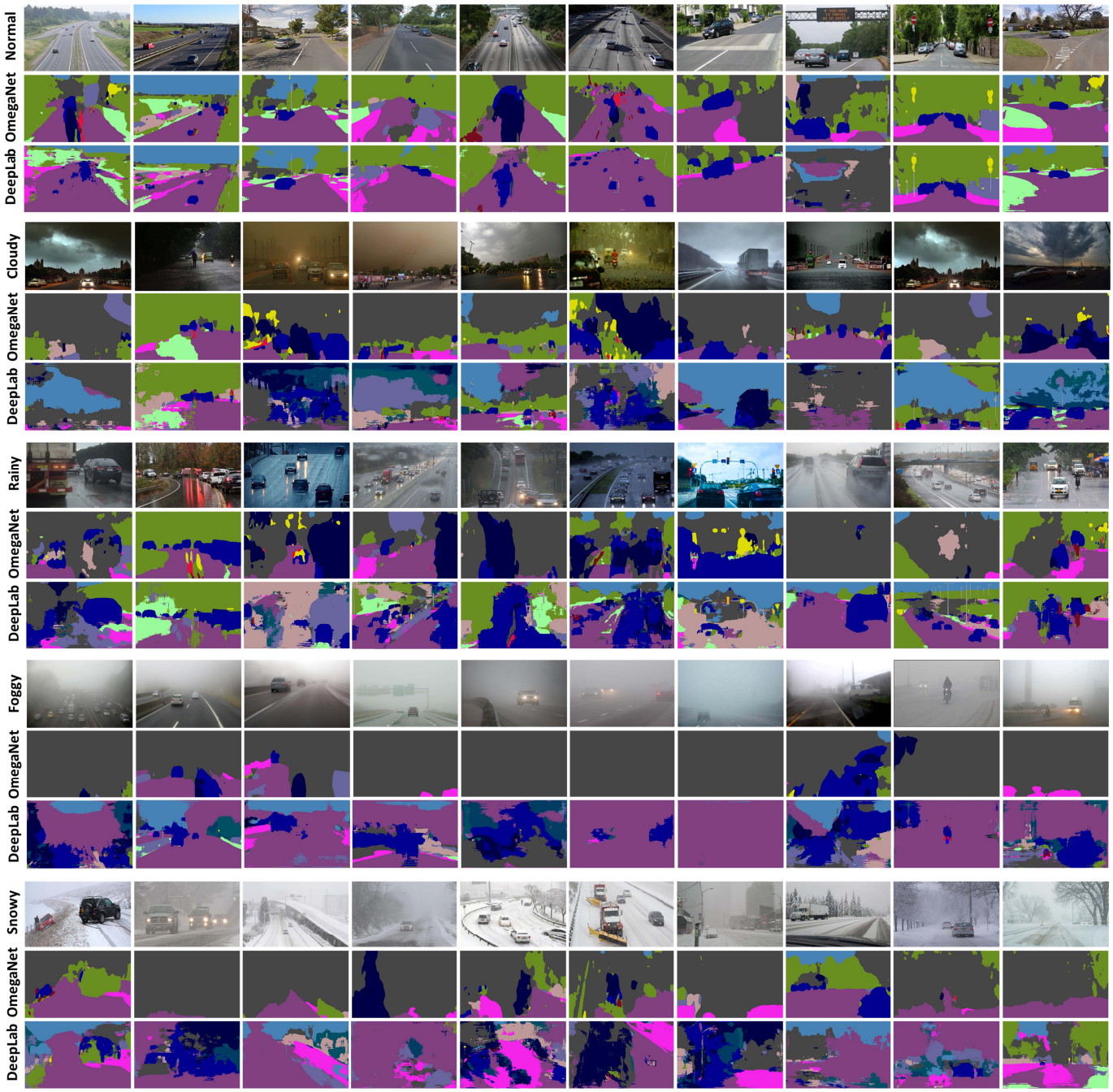


Fig. 5. Visual segmentation results of various deep segmentation models over input images having variable weather conditions such as snow and fog. The segmented maps of DeepLab and OmegaNet are not trustworthy for an autonomous vehicle's scene understanding.

that are unaffordable for their adoption in real vehicular on-board hardware. Despite this noted issue, more generally there are some scenes encountered while driving which have overlap and occlusion during consecutive frames, paralysing the frame-based segmentation for scene understanding. Video-based segmentation is a contemporary option in this regard, which should ensure faster processing and a better practicality for AD applications.

3) Object's Predicted Locations Segmentation: A significantly vibrant research activity can be lately noted around the estimation of the future location of pedestrians and other moving objects in the scene, such as vehicles [60]. Notwithstanding its highly challenging nature, the task of future location estimation assists decision making of autonomous vehicle, providing estimated future trajectories of persons and vehicles. Unfortunately, research revolving on segmenting future locations is scarce and to the best of our knowledge there is not a single research segmenting or drawing segmentation maps of pedestrian or other objects' future locations. This area is very challenging though, but not far to be achieved for scene understanding. Recently, many methods [129], [130] have achieved accurate bounding-boxes prediction of pedestrians

for upcoming 10 to 15 frames. These methods can be considered as the baseline for future research in this valuable direction.

4) Hybrid Methods and Multi-Modalities: Besides the broader coverage of RGB data generated by vision sensors, there are some other modalities [131] and sensors with quite informative patterns and points for scene analysis and understanding. For instance, point clouds [132], [133], meshes [134] and depth data [135] together with RGB data can generate an increased 3D scene understanding for an autonomous vehicle [136]. These data are generated from various sensors, including LiDAR among many other options. Hybrid models are widely used in many domains [137], [138] with successful results in terms of vulnerability and can be implemented in ITS domain as well. As vehicles are equipped with more sensors, we envision many opportunities for research on multi-modal information fusion, further stimulated by other non-embarked sources of related information (e.g. floating car data – wherein cell phones of drivers and passengers act as additional traffic probes – or social network data).

5) Active and Incremental Learning: Active learning [139] in machine learning refers to self-adaptability and learning of a model with respect to time and new data encountered during testing phase even after its deployment stage [140]. In real-world environments, dynamic scenes with rarely occurring living species or objects such as kangaroo or a self-engineered dump and cargo truck may be encountered by a vehicle, which may rely on AI's model decision for further actions such as applying brake or increasing acceleration. Thus, a scene understanding AI based mechanism should interactively allow processing queries of every type of data and its structures in the form of unlabeled data instances labeled by a human annotator during the process, involving human in the training loop [141]. There are different types of active learning techniques, such as membership query synthesis [142], where synthetic data is generated and the parameters of synthetic data can be tuned [143] based on structure of objects, derived from base species of the dataset. On the other hand, the capability of segmentation models to update their captured knowledge with new data in an incremental fashion is a key for their sustainability and continuous improvement. We foresee that these two capabilities of segmentation models for scene understanding will grow in importance in prospective studies.

6) Complex Driving Scenes Understanding: Semantic segmentation with applications to scene understanding primarily focuses on objects in a single category without any consideration to the importance of their location. For instance, a pedestrian walking through a sidewalk is classified simply as a pedestrian. There are some disadvantages associated to this approach: the extra time involved by an algorithm to verify its location; and let the vehicle decide actions, there is no specific safety levels of pedestrians (relate-able to cyclist and other objects), treating all objects as belonging to a similar safety level. Therefore, for complex driving scenes with abundant human subjects, there is a need for priority-driven systems to segment the pedestrians on vehicular lanes in a different category, and conversely, for the pedestrian with huge distance or ones on side walk. A baseline research dealing with this problem recently introduced a pedestrian location perception network with location inference of each semantic map corresponding to the human [56]. This work can be advanced in terms of more objects identified in scenes characterized by a higher complexity and diversity.

7) Adverse Weather Conditions: When operative in real-world environments, autonomous vehicles may encounter adverse weather conditions such as snow, fog, rain or dark areas, among other phenomena [144], [145]. Existing models are highly accurate for normal cases with sufficient illumination and other favorable conditions. However, models need to be adaptable to non-favorable weather scenarios. For instance, a dataset for night-time segmentation is introduced by Xin *et al.* in [146]. Furthermore, preprocessing techniques for haze [147] and fog [148] removal ensure effective semantic segmentation. But at the same time, if deep segmentation models are designed with built-in capabilities to account for weather-related uncertainties, or they prove to be effective in such cases, would decrease the computation time required for the aforementioned preprocessing steps. Some representative results of existing models over weather uncertainties are tested and reported in Figure 5, whereas a baseline research for scene understanding has developed a deep model and a Foggy Cityscapes dataset [149]. The segmentation maps generated by these models clearly outline a long road ahead in this direction. The current models seem to have insufficient generalization potentials towards challenging scenarios such as rainy environment, snow, and cloudy scenarios. Despite the presence of some challenging datasets in adverse weather conditions such as Fog [150], [151], night time and dark scenarios [152], [153], wild [154], *etc.*, the current methods still lack focusing on end-to-end deep models to handle complex weather scenarios effectively. There also exists some generalized datasets with multiple challenges [116], [155], but the amount of data labelled for semantic segmentation in most of these datasets are very limited *i.e.,* number of annotated instances ranging from 40 [151] to maximum 4006 [155] samples.

Utilization of advanced driving simulators such as VituoCity [156] to create photo-realistic synthetic dataset without needing expensive and high-risk driving in real-world is also among the current approaches to compensate experiments in adverse weather conditions.

8) Events-Based Scene Understanding: So far, scene understanding has been primarily approached by using segmentation techniques. Nonetheless, the focus can be diverted towards higher levels of vehicular cognition, such as events based scene understanding [157]. For instance, analyzing the events for scene parsing is a promising direction, where surrounding events such as bicyclist on the vehicle lane, pedestrian crossing the road, among many other common events can better support and favor more informed decision making of autonomous vehicles [158]. The main point here is to not rely only on segmentation for scene understanding, but rather to explore other metrics and to discover relationships between identified objects over space and time [159]. It is our belief that this augmented contextual awareness will be a major breakthrough towards the accountability of decisions made by autonomous vehicles.

9) Replacing CNNs With Vision Transformers: Dense prediction models, such as semantic segmentation and saliency detection, are mostly inspired by convolutional architectures. Particularly, backbones of semantic segmentation methods mainly rely on convolutional operations. It is true that these networks progressively downsample input images and acquire features at multiple scales, thus allowing for increased receptive fields. These mechanisms for feature refining, *i.e.,* for transitioning from low-level to high-level descriptors, are computationally complex and have certain limitations for many computer vision tasks, particularly for dense prediction tasks. For instance, the granularity of the features, as well as their resolution, are lost gradually as the layers go deeper and deeper, by producing inadequate representations for subsequent decoder layers, and by loosing information that cannot be recovered during the decoding procedure. Training at higher input resolutions demands higher computational budget, whereas the use of dilated convolutions increases receptive fields quickly without downsampling. Other similar techniques can be applied to mitigate the loss of feature granularity. Unfortunately, such techniques still suffer from bottlenecks due to the involvement of convolutional operations over the hierarchical neural structure of the model.

In contrast, transformers (as encoders) have better image representation capabilities [160], [161], which mainly hinge on representing images as *bag-of-words*, and passing them through various transformer layers to extract features at several resolutions. Then, they progressively integrate these multi-resolution representations to finally attain the concerned dense prediction task. When trained over large-scale datasets, vision transformers [162], [163], [164], [165], [166], [167], [168], [169] perform well for dense prediction tasks. For instance, Ranftl *et al.* [162] establish an unprecedented state-of-the-art level of performance by introducing vision transformers in a semantic segmentation domain. A similar approach is observable for the saliency detection domain, where the authors in [163] applied vision transformers with multi-level tokens fusion and a new token upsampling strategy based on transformers. Liu *et al.*, [165] introduced a transformer-based weakly supervised semantic segmentation method named WegFormer, which encapsulated three different components to generate high-quality segmentation masks. Their presented WegFormer first generates attention maps using deep taylor decomposition (DTD) and then used a soft erasing mechanism to smooth computed attention maps. Finaly, they have filtered the noisy activation maps using their proposed efficient potential object mining strategy. Ruiping *et al.*, [166] presented knowledge distillation driven transformer for efficient semantic segmentation of road scenes. They have retrained a shallow transformer by transferring the learned knowledge from large transformer network trained on large volume of image data. The knowledge distillation strategy allowed their method to achieve the same level of segmentation performance and faster inference time due to reduced computational complexity. Lin *et al.*, [168] proposed a multi-scale transformer for efficient semantic segmentation, which extracts multi-level features from an image and then aggregates the extracted features using a feature selection technique. The aggregated features are then used to determine the salient regions of the given image, resulting a fine quality semantic segmentation. So far, these methods have achieved unrivaled performance levels in these specific domains, unleashing manifold future research directions and opportunities for semantic segmentation tasks.

10) Towards More Accurate and Efficient Semantic Segmentation Methods for AD: The qualitative performance of currently employed semantic segmentation techniques is shown in Table VI, where we notice that only a few methods are able to balance the trade-off between accuracy and inference latency of their model. The experimental results reported for these methods indicate that they require further work to alleviate their computational burden while maintaining their unparalleled performance. Furthermore, we test most well-known semantic segmentation models in a few challenging scenarios, as reported in Figure 5. We have found that these models should also be evaluated in terms of knowledge transferability and generalization accross different datasets [89]. Furthermore, the time complexity reported in Table VI suggests that some of these methods are functional in real time when deployed on GPU devices. In any case, the focus of semantic segmentation methods should also be diverted towards computational complexity, given the stringently limited computational resources available in today's AD in-vehicle telematics.

VI. Concluding Remarks and Outlook

Vision sensors' data are a key component of autonomous vehicles, playing a significant role in an autonomous vehicle's decision making. Vision sensory data are analyzed using Computational Intelligence techniques for effective outputs such as sign board detection, drivable area selection, and traffic lights perception. In doing so, an autonomous vehicle senses the surroundings using vision sensory data. Segmentation extracts pixel values of various objects inside an input image and individuates them from one another using distinct colors. The segmentation of various objects into their respective classes helps dramatically in parsing scene information for the vehicle. Although complementary options can be found to derive data from other sensors for decision making, vision sensors have undoubtedly a major role in the current vehicular panorama.

Segmentation for scene understanding of autonomous vehicles has been in play for many years, but a consolidated, summarized analysis is absent from the existing literature. In this survey we have discussed on the strengths of existing segmentation methods in clear environments and their weaknesses when facing challenging scenarios. Our main conclusion is that the scene understanding literature has not achieved perfection yet, as many limitations remain in the current methods that we have thoroughly covered in our review, followed by relevant suggestions and outlooks in a detailed manner. We have covered baseline works dealing with deep learning models, their hierarchy for segmentation tasks and the challenges associated to each model category. Furthermore, performance evaluation strategies suited for segmentation models, special loss functions, and datasets widely used in AD domain have been also tackled in depth. We have

rounded up the work by exposing open challenges for scene understanding, together with future research directions with stimulating baseline references from the recent literature.

On a closing note, it is undeniable that experts from the ITS community are continuously struggling towards better scene understanding strategies to utilize the vision sensors' data effectively. Mainstream research is gravitated towards improving the model's accuracy through the capabilities of its neural layers. However, there exist other challenges to be covered in order to achieve reliable, trustworthy and safe AD. Challenges from the scene understanding perspective demand robust models with prioritization levels for segmented objects, coarse-structure information processing capabilities, and risk categorization. Furthermore, current deep segmentation models are confied to handle a single information modality, while recently point cloud data [133], [170] have been studied extensively for complex tasks related to AD. These are open opportunities to utilize multi-modal data such as 3D LiDAR [171] and vision sensors, and to transcend from single deep neural network to more elaborated fusion models, capable of accomplishing complicated yet more informative learning tasks for autonomous vehicles. These opportunities (if well and timely leveraged) can advance the ITS research and bring scene segmentation to a new level, where driver-less vehicles can be deployed in real-world environments and support safer and reliable travel services.

REFERENCES

[1] W. Wei, R. Yang, H. Gu, W. Zhao, C. Chen, and S. Wan, "Multi-objective optimization for resource allocation in vehicular cloud computing networks," *IEEE Trans. Intell. Transp. Syst.*, early access, Aug. 3, 2021, doi: 10.1109/TITS.2021.3091321.

[2] Q. Zou, Q. Sun, L. Chen, B. Nie, and Q. Li, "A comparative analysis of LiDAR SLAM-based indoor navigation for autonomous vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 6907–6921, Jul. 2022.

[3] C. Chen, L. Liu, S. Wan, X. Hui, and Q. Pei, "Data dissemination for industry 4.0 applications in Internet of vehicles based on short-term traffic prediction," *ACM Trans. Internet Technol.*, vol. 22, no. 1, pp. 1–18, Feb. 2022.

[4] M. Usman, M. A. Jan, and A. Jolfaei, "SPEED: A deep learning assisted privacy-preserved framework for intelligent transportation systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 7, pp. 4376–4384, Jul. 2021.

[5] S. U. Khan, T. Hussain, A. Ullah, and S. W. Baik, "Deep-ReID: Deep features and autoencoder assisted image patching strategy for person re-identification in smart cities surveillance," *Multimedia Tools Appl.*, early access, pp. 1–22, Jan. 2021.

[6] P. Dhawankar, P. Agrawal, B. Abderezzak, O. Kaiwartya, K. Busawon, and M. S. Raboacă, "Design and numerical implementation of V2X control architecture for autonomous driving vehicles," *Mathematics*, vol. 9, no. 14, p. 1696, 2021.

[7] O. Kaiwartya *et al.*, "Internet of vehicles: Motivation, layered architecture, network model, challenges, and future aspects," *IEEE Access*, vol. 4, pp. 5356–5373, 2016.

[8] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[9] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Oct. 2016.

[10] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.

[11] M. Kurdej, J. Moras, V. Cherfaoui, and P. Bonnifait, "Map-aided evidential grids for driving scene understanding," *IEEE Intell. Transp. Syst. Mag.*, vol. 7, no. 1, pp. 30–41, Jan. 2015.

[12] N. Dhanachandra, Y. J. Chanu, and K. M. Singh, "A new hybrid image segmentation approach using clustering and black hole algorithm," *Comput. Intell.*, Mar. 2020.

[13] P. Meletis, "Towards holistic scene understanding: Semantic segmentation and beyond," 2022, *arXiv:2201.07734*.

[14] S. Khan, H. Rahmani, S. A. A. Shah, and M. Bennamoun, "A guide to convolutional neural networks for computer vision," *Synth. Lect. Comput. Vis.*, vol. 8, no. 1, pp. 1–207, 2018.

[15] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1520–1528.

[16] S. Wan, S. Ding, and C. Chen, "Edge computing enabled video segmentation for real-time traffic monitoring in internet of vehicles," *Pattern Recognit.*, vol. 121, Jan. 2022, Art. no. 108146.

[17] H. Shi, M. Hayat, and J. Cai, "Transformer scale gate for semantic segmentation," 2022, *arXiv:2205.07056*.

[18] D. Bogdoll, E. Eisen, M. Nitsche, C. Scheib, and J. M. Zöllner, "Multimodal detection of unknown objects on roads for autonomous driving," 2022, *arXiv:2205.01414*.

[19] F. Nesti, G. Rossolini, S. Nair, A. Biondi, and G. Buttazzo, "Evaluating the robustness of semantic segmentation for autonomous driving against real-world adversarial patch attacks," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 2280–2289.

[20] G. Rossolini, F. Nesti, G. D'Amico, S. Nair, A. Biondi, and G. Buttazzo, "On the real-world adversarial robustness of real-time semantic segmentation models for autonomous driving," 2022, *arXiv:2201.01850*.

[21] O. Natan and J. Miura, "Towards compact autonomous driving perception with balanced learning and multi-sensor fusion," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 16249–16266, Sep. 2022.

[22] Y. Zhao, W. Tian, and H. Cheng, "Pyramid Bayesian method for model uncertainty evaluation of semantic segmentation in autonomous driving," *Automot. Innov.*, vol. 5, no. 1, pp. 70–78, 2022.

[23] O. Natan and J. Miura, "Semantic segmentation and depth estimation with RGB and DVS sensor fusion for multi-view driving perception," in *Proc. Asian Conf. Pattern Recognit.* Cham, Switzerland: Springer, 2022, pp. 352–365.

[24] Y. B. Can, A. Liniger, O. Unal, D. Paudel, and L. Van Gool, "Understanding bird's-eye view of road semantics using an onboard camera," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 3302–3309, Apr. 2022.

[25] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50× fewer parameters and <0.5 MB model size," 2016, *arXiv:1602.07360*.

[26] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "ENet: A deep neural network architecture for real-time semantic segmentation," 2016, *arXiv:1606.02147*.

[27] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiseNet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 325–341.

[28] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "ICNet for real-time semantic segmentation on high-resolution images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 405–420.

[29] J.-R. Xue, J.-W. Fang, and P. Zhang, "A survey of scene understanding by event reasoning in autonomous driving," *Int. J. Autom. Comput.*, vol. 15, no. 3, pp. 249–266, 2018.

[30] M. Naseer, S. Khan, and F. Porikli, "Indoor scene understanding in 2.5/3D for autonomous agents: A survey," *IEEE Access*, vol. 7, pp. 1859–1887, 2019.

[31] Z. Huang, C. Lv, Y. Xing, and J. Wu, "Multi-modal sensor fusion-based deep neural network for end-to-end autonomous driving with scene understanding," *IEEE Sensors J.*, vol. 21, no. 10, pp. 11781–11790, May 2021.

[32] Y. Zhang, O. Morel, M. Blanchon, R. Seulin, M. Rastgoo, and D. Sidibé, "Exploration of deep learning-based multimodal fusion for semantic road scene segmentation," in *Proc. 14th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, 2019, pp. 336–343.

[33] M. Schon, M. Buchholz, and K. Dietmayer, "MGNet: Monocular geometric scene understanding for autonomous driving," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15804–15815.

[34] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "DenseASPP for semantic segmentation in street scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3684–3692.

[35] M. P. Muresan, I. Giosan, and S. Nedevschi, "Stabilization and validation of 3D object position using multimodal sensor fusion and semantic segmentation," *Sensors*, vol. 20, no. 4, p. 1110, Feb. 2020.

[36] J. Fei, W. Chen, P. Heidenreich, S. Wirges, and C. Stiller, "SemanticVoxels: Sequential fusion for 3D pedestrian detection using LiDAR point cloud and semantic segmentation," 2020, *arXiv:2009.12276*.

[37] F. Langer, A. Milioto, A. Haag, J. Behley, and C. Stachniss, "Domain transfer for semantic segmentation of LiDAR data using deep neural networks," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 8263–8270.

[38] F. Duerr, H. Weigel, M. Maehlisch, and J. Beyerer, "Iterative deep fusion for 3D semantic segmentation," in *Proc. 4th IEEE Int. Conf. Robotic Comput. (IRC)*, Nov. 2020, pp. 391–397.

[39] Z. Zhuang, R. Li, K. Jia, Q. Wang, Y. Li, and M. Tan, "Perception-aware multi-sensor fusion for 3D LiDAR semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 16280–16290.

[40] L. Zhao, H. Zhou, X. Zhu, X. Song, H. Li, and W. Tao, "LIF-Seg: LiDAR and camera image fusion for 3D LiDAR semantic segmentation," 2021, *arXiv:2108.07511*.

[41] Y. Wu, "DenseFuseNet: Improve 3D semantic segmentation in the context of autonomous driving with dense correspondence," in *Proc. IEEE Int. Conf. Consum. Electron. Comput. Eng. (ICCECE)*, Jan. 2021, pp. 259–270.

[42] L. Caltagirone, M. Bellone, L. Svensson, M. Wahde, and R. Sell, "LiDAR–camera semi-supervised learning for semantic segmentation," *Sensors*, vol. 21, no. 14, p. 4813, Jul. 2021.

[43] Z. Wang *et al.*, "Multi-stage fusion for multi-class 3D LiDAR detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 3120–3128.

[44] F. Duerr, H. Weigel, and J. Beyerer, "Decoupled iterative deep sensor fusion for 3D semantic segmentation," *Int. J. Semantic Comput.*, vol. 15, no. 3, pp. 293–312, Sep. 2021.

[45] V. Hassija, V. Gupta, S. Garg, and V. Chamola, "Traffic jam probability estimation based on blockchain and deep neural networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 7, pp. 3919–3928, Jul. 2021.

[46] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.

[47] J. Behley *et al.*, "SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9297–9307.

[48] B. Gao, Y. Pan, C. Li, S. Geng, and H. Zhao, "Are we hungry for 3D LiDAR data for semantic segmentation? A survey of datasets and methods," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 6063–6081, Jul. 2022.

[49] M. Hofmarcher, T. Unterthiner, J. Arjona-Medina, G. Klambauer, S. Hochreiter, and B. Nessler, "Visual scene understanding for autonomous driving using semantic segmentation," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Berlin, Germany: Springer, 2019, pp. 285–296.

[50] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.

[51] K. Chen *et al.*, "MVLidarNet: Real-time multi-class scene understanding for autonomous driving using multiple views," 2020, *arXiv:2006.05518*.

[52] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, 2013.

[53] H. Caesar *et al.*, "NuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11621–11631.

[54] M. Kalweit, G. Kalweit, M. Werling, and J. Boedecker, "Deep surrogate Q-learning for autonomous driving," 2020, *arXiv:2010.11278*.

[55] R. Krajewski, J. Bock, L. Kloeker, and L. Eckstein, "The high D dataset: A drone dataset of naturalistic vehicle trajectories on German highways for validation of highly automated driving systems," in *Proc. 21st Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2018, pp. 2118–2125.

[56] Y. Zhao, M. Qi, X. Li, Y. Meng, Y. Yu, and Y. Dong, "P-LPN: Towards real time pedestrian location perception in complex driving scenes," *IEEE Access*, vol. 8, pp. 54730–54740, 2020.

[57] F. Tosi *et al.*, "Distilled semantics for comprehensive scene understanding from videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4654–4665.

[58] L. Sless, B. E. Shlomo, G. Cohen, and S. Oron, "Road scene understanding by occupancy grid learning from sparse radar clusters using semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1–9.

[59] D. Neven, B. De Brabandere, S. Georgoulis, M. Proesmans, and L. Van Gool, "Fast scene understanding for autonomous driving," 2017, *arXiv:1708.02550*.

[60] A. Hu, F. Cotter, N. Mohan, C. Gurau, and A. Kendall, "Probabilistic future prediction for video scene understanding," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 767–785.

[61] X. Huang *et al.*, "The ApolloScape dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 954–960.

[62] G. Neuhold, T. Ollmann, S. R. Bulo, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4990–4999.

[63] F. Yu *et al.*, "BDD100K: A diverse driving dataset for heterogeneous multitask learning," 2018, *arXiv:1805.04687*.

[64] J. Baek *et al.*, "Scene understanding networks for autonomous driving based on around view monitoring system," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 961–968.

[65] N. Dvornik, K. Shmelkov, J. Mairal, and C. Schmid, "BlitzNet: A real-time deep network for scene understanding," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4154–4162.

[66] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.

[67] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[68] M. Everingham and J. Winn, "The Pascal visual object classes challenge 2012 (VOC2012) development kit," in *Pattern Analysis, Statistical Modelling and Computational Learning*, vol. 8. 2011.

[69] V. Ramanishka, Y.-T. Chen, T. Misu, and K. Saenko, "Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7699–7707.

[70] X. Wang, H. Ma, and S. You, "Deep clustering for weakly-supervised semantic segmentation in autonomous driving scenes," *Neurocomputing*, vol. 381, pp. 20–28, Mar. 2020.

[71] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern Recognit. Lett.*, vol. 30, no. 2, pp. 88–97, 2009.

[72] L. Chen, W. Zhan, W. Tian, Y. He, and Q. Zou, "Deep integration: A multi-label architecture for road scene recognition," *IEEE Trans. Image Process.*, vol. 28, no. 10, pp. 4883–4898, May 2019.

[73] P. Zhang, W. Liu, Y. Lei, H. Wang, and H. Lu, "RAPNet: Residual atrous pyramid network for importance-aware street scene parsing," *IEEE Trans. Image Process.*, vol. 29, pp. 5010–5021, 2020.

[74] B. Baheti, S. Innani, S. Gajre, and S. Talbar, "Eff-UNet: A novel architecture for semantic segmentation in unstructured environment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 358–359.

[75] C. H. Bahnsen and T. B. Moeslund, "Rain removal in traffic surveillance: Does it matter?" *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 8, pp. 2802–2819, Aug. 2019.

[76] D. Gangodkar, P. Kumar, and A. Mittal, "Robust segmentation of moving vehicles under complex outdoor conditions," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 4, pp. 1738–1752, Dec. 2012.

[77] S. Di *et al.*, "Rainy night scene understanding with near scene semantic adaptation," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 3, pp. 1594–1602, Mar. 2021.

[78] K. Muhammad, A. Ullah, J. Lloret, J. D. Ser, and V. H. C. de Albuquerque, "Deep learning for safe autonomous driving: Current challenges and future directions," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 7, pp. 4316–4336, Jul. 2021.

[79] D. González, J. Pérez, V. Milanés, and F. Nashashibi, "A review of motion planning techniques for automated vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 4, pp. 1135–1145, Apr. 2016.

[80] B. Paden, M. Čáp, S. Z. Yong, D. Yershov, and E. Frazzoli, "A survey of motion planning and control techniques for self-driving urban vehicles," *IEEE Trans. Intell. Veh.*, vol. 1, no. 1, pp. 33–55, Mar. 2016.

[81] G. Bresson, Z. Alsayed, L. Yu, and S. Glaser, "Simultaneous localization and mapping: A survey of current trends in autonomous driving," *IEEE Trans. Intell. Veh.*, vol. 2, no. 3, pp. 194–220, Sep. 2017.

[82] J. Wang, J. Liu, and N. Kato, "Networking and communications in autonomous driving: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 2, pp. 1243–1274, 2nd Quart., 2019.

[83] S. Kuutti, S. Fallah, K. Katsaros, M. Dianati, F. Mccullough, and A. Mouzakitis, "A survey of the state-of-the-art localization techniques and their potentials for autonomous vehicle applications," *IEEE Internet Things J.*, vol. 5, no. 2, pp. 829–846, Apr. 2018.

[84] A. Rasouli and J. K. Tsotsos, "Autonomous vehicles that interact with pedestrians: A survey of theory and practice," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 3, pp. 900–918, Mar. 2020.

[85] M. Gerstmair, A. Melzer, A. Onic, and M. Huemer, "On the safe road toward autonomous driving: Phase noise monitoring in radar sensors for functional safety compliance," *IEEE Signal Process. Mag.*, vol. 36, no. 5, pp. 60–70, Sep. 2019.

[86] I. Bilik, O. Longman, S. Villeval, and J. Tabrikian, "The rise of radar for autonomous vehicles: Signal processing solutions and future research directions," *IEEE Signal Process. Mag.*, vol. 36, no. 5, pp. 20–31, Sep. 2019.

[87] E. Arnold, O. Y. Al-Jarrah, M. Dianati, S. Fallah, D. Oxtoby, and A. Mouzakitis, "A survey on 3D object detection methods for autonomous driving applications," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 10, pp. 3782–3795, Oct. 2019.

[88] J. Janai, F. Gü ney, A. Behl, and A. Geiger, "Computer vision for autonomous vehicles: Problems, datasets and state of the art," *Found. Trends Comput. Graph. Vis.*, vol. 12, nos. 1–3, pp. 1–308, 2020.

[89] T. Hussain, S. Anwar, A. Ullah, K. Muhammad, and S. W. Baik, "Densely deformable efficient salient object detection network," 2021, *arXiv:2102.06407*.

[90] M. A. Rahman and Y. Wang, "Optimizing intersection-over-union in deep neural networks for image segmentation," in *Proc. Int. Symp. Vis. Comput.* Cham, Switzerland: Springer, 2016, pp. 234–244.

[91] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 658–666.

[92] S. Jadon, "A survey of loss functions for semantic segmentation," in *Proc. IEEE Conf. Comput. Intell. Bioinf. Comput. Biol. (CIBCB)*, Oct. 2020, pp. 1–7.

[93] Y. S. Aurelio, G. M. de Almeida, C. L. de Castro, and A. P. Braga, "Learning from imbalanced data sets with weighted cross-entropy function," *Neural Process. Lett.*, vol. 50, no. 2, pp. 1937–1949, Oct. 2019.

[94] Y. Ho and S. Wookey, "The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling," *IEEE Access*, vol. 8, pp. 4806–4813, 2020.

[95] S. Pan *et al.*, "Diagnostic model of coronary microvascular disease combined with full convolution deep network with balanced cross-entropy cost function," *IEEE Access*, vol. 7, pp. 177997–178006, 2019.

[96] K. Hu *et al.*, "Retinal vessel segmentation of color fundus images using multiscale convolutional neural network with an improved cross-entropy loss function," *Neurocomputing*, vol. 309, pp. 179–191, Oct. 2018.

[97] N. Abraham and N. M. Khan, "A novel focal Tversky loss function with improved attention U-Net for lesion segmentation," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 683–687.

[98] H. Caesar, J. Uijlings, and V. Ferrari, "Region-based semantic segmentation with end-to-end training," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 381–397.

[99] M.-P. Dubuisson and A. K. Jain, "A modified Hausdorff distance for object matching," in *Proc. 12th IAPR Int. Conf. Pattern Recognit. (ICPR)*, vol. 1. Oct. 1994, pp. 566–568.

[100] N. Aspert, D. Santa-Cruz, and T. Ebrahimi, "MESH: Measuring errors between surfaces using the Hausdorff distance," in *Proc. IEEE Int. Conf. Multimedia Expo*, Aug. 2002, pp. 705–708.

[101] S. Wonneberger and S. Balci, "Reliable validation of highly automated driving functions by increasing the virtualization level of high performance computing platforms and smart sensors," *ELIV 2021*, 2021.

[102] H. Li, P. Xiong, H. Fan, and J. Sun, "DFANet: Deep feature aggregation for real-time semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9522–9531.

[103] N. Khan, A. Ullah, I. U. Haq, V. G. Menon, and S. W. Baik, "SD-Net: Understanding overcrowded scenes in real-time via an efficient dilated convolutional neural network," *J. Real-Time Image Process.*, vol. 18, no. 5, pp. 1729–1743, 2020.

[104] N. Dilshad, J. Hwang, J. Song, and N. Sung, "Applications and challenges in video surveillance via drone: A brief survey," in *Proc. Int. Conf. Inf. Commun. Technol. Converg. (ICTC)*, Oct. 2020, pp. 728–732.

[105] K. Muhammad, T. Hussain, J. J. P. C. Rodrigues, P. Bellavista, A. R. L. de Macedo, and V. H. C. de Albuquerque, "Efficient and privacy preserving video transmission in 5G-enabled IoT surveillance networks: Current challenges and future directions," *IEEE Netw.*, vol. 35, no. 2, pp. 26–33, Mar. 2021.

[106] V. Hassija, V. Chamola, V. Saxena, D. Jain, P. Goyal, and B. Sikdar, "A survey on IoT security: Application areas, security threats, and solution architectures," *IEEE Access*, vol. 7, pp. 82721–82743, 2019.

[107] R. Mohan and A. Valada, "Efficientps: Efficient panoptic segmentation," *Int. J. Comput. Vis.*, vol. 129, no. 5, pp. 1551–1579, 2021.

[108] S. A. Kamran and A. S. Sabbir, "Efficient yet deep convolutional neural networks for semantic segmentation," in *Proc. Int. Symp. Adv. Intell. Informat. (SAIN)*, Aug. 2018, pp. 123–130.

[109] S. Choi, S. Jung, H. Yun, J. Kim, S. Kim, and J. Choo, "RobustNet: Improving domain generalization in urban-scene segmentation via instance selective whitening," 2021, *arXiv:2103.15597*.

[110] P.-R. Chen, H.-M. Hang, S.-W. Chan, and J.-J. Lin, "DSNet: An efficient CNN for road scene segmentation," *APSIPA Trans. Signal Inf. Process.*, vol. 9, no. 1, pp. 1–14, 2020.

[111] Q. Lv, X. Sun, C. Chen, J. Dong, and H. Zhou, "Parallel complement network for real-time semantic segmentation of road scenes," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 5, pp. 4432–4444, May 2022.

[112] J. Fu, J. Liu, J. Jiang, Y. Li, Y. Bao, and H. Lu, "Scene segmentation with dual relation-aware attention network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 6, pp. 2547–2560, Jun. 2021.

[113] X. Roynard, J.-E. Deschaud, and F. Goulette, "Paris-Lille-3D: A large and high-quality ground-truth urban point cloud dataset for automatic segmentation and classification," *Int. J. Robot. Res.*, vol. 37, no. 6, pp. 545–557, May 2018.

[114] J. Behley, V. Steinhage, and A. B. Cremers, "Performance of histogram descriptors for the classification of 3D laser range data in urban environments," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2012, pp. 4391–4398.

[115] X. Huang, P. Wang, X. Cheng, D. Zhou, Q. Geng, and R. Yang, "The ApolloScape open dataset for autonomous driving and its application," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2702–2719, Oct. 2020.

[116] F. Yu *et al.*, "BDD100K: A diverse driving dataset for heterogeneous multitask learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2636–2645.

[117] X. Wu, Z. Wu, H. Guo, L. Ju, and S. Wang, "DANNet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation," 2021, *arXiv:2104.10834*.

[118] M. Siam, H. Mahgoub, M. Zahran, S. Yogamani, M. Jagersand, and A. El-Sallab, "MODNet: Motion and appearance based moving object detection network for autonomous driving," in *Proc. 21st Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2018, pp. 2859–2864.

[119] F. Arnez, H. Espinoza, A. Radermacher, and F. Terrier, "A comparison of uncertainty estimation approaches in deep learning components for autonomous vehicle applications," 2020, *arXiv:2006.15172*.

[120] T. Cortinhal, G. Tzelepis, and E. E. Aksoy, "SalsaNext: Fast, uncertainty-aware semantic segmentation of LiDAR point clouds for autonomous driving," 2020, *arXiv:2003.03653*.

[121] R. Michelmore, M. Kwiatkowska, and Y. Gal, "Evaluating uncertainty quantification in end-to-end autonomous driving control," 2018, *arXiv:1811.06817*.

[122] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," 2018, *arXiv:1806.01768*.

[123] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/2650d6089a6d640c5e85b2b88265dc2b-Paper.pdf

[124] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1050–1059.

[125] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2017, pp. 6405–6416.

[126] M. Abdar *et al.*, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," 2020, *arXiv:2011.06225*.

[127] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3319–3328.

[128] A. B. Arrieta *et al.*, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.

[129] K. Mangalam, E. Adeli, K.-H. Lee, A. Gaidon, and J. C. Niebles, "Disentangling human dynamics for pedestrian locomotion forecasting with noisy supervision," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2784–2793.

[130] O. Styles, T. Guha, and V. Sanchez, "Multiple object forecasting: Predicting future object locations in diverse environments," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 690–699.

[131] S. P. Sotiroudis, P. Sarigiannidis, S. K. Goudos, and K. Siakavara, "Fusing diverse input modalities for path loss prediction: A deep learning approach," *IEEE Access*, vol. 9, pp. 30441–30451, 2021.

[132] F. Zhang *et al.*, "Instance segmentation of LiDAR point clouds," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 9448–9455.

[133] Y. Cui *et al.*, "Deep learning for image and point cloud fusion in autonomous driving: A review," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 2, pp. 722–739, Feb. 2022.

[134] S. Liu, J.-L. He, and S.-H. Liao, "Automatic detection of anatomical landmarks on geometric mesh data using deep semantic segmentation," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2020, pp. 1–6.

[135] A. Ahmed, A. Jalal, and K. Kim, "RGB-D images for object segmentation, localization and recognition in indoor scenes using feature descriptor and Hough voting," in *Proc. 17th Int. Bhurban Conf. Appl. Sci. Technol. (IBCAST)*, Jan. 2020, pp. 290–295.

[136] G. Krispel, M. Opitz, G. Waltner, H. Possegger, and H. Bischof, "FuseSeg: LiDAR point cloud segmentation fusing multi-modal data," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1874–1883.

[137] A. Ur-Rehman, I. Gondal, J. Kamruzzaman, and A. Jolfaei, "Vulnerability modelling for hybrid industrial control system networks," *J. Grid Comput.*, vol. 18, no. 4, pp. 863–878, Dec. 2020.

[138] A. Ur-Rehman, I. Gondal, J. Kamruzzuman, and A. Jolfaei, "Vulnerability modelling for hybrid IT systems," in *Proc. IEEE Int. Conf. Ind. Technol. (ICIT)*, Feb. 2019, pp. 1186–1191.

[139] P. Ren *et al.*, "A survey of deep active learning," 2020, *arXiv:2009.00236*.

[140] S. Budd, E. C. Robinson, and B. Kainz, "A survey on active learning and human-in-the-loop deep learning for medical image analysis," *Med. Image Anal.*, vol. 71, Jul. 2021, Art. no. 102062.

[141] A. Khan *et al.*, "PMAL: A proxy model active learning approach for vision based industrial applications," *ACM Trans. Multimedia Comput. Commun. Appl.*, Jun. 2022.

[142] O. Kelner, "Learning halfspaces with membership queries," 2020, *arXiv:2012.10985*.

[143] M. Ahmed *et al.*, "AAQAL: A machine learning-based tool for performance optimization of parallel SPMV computations using block CSR," *Appl. Sci.*, vol. 12, no. 14, p. 7073, Jul. 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/14/7073

[144] J. Vertens, J. Zürn, and W. Burgard, "HeatNet: Bridging the day-night domain gap in semantic segmentation with thermal images," 2020, *arXiv:2003.04645*.

[145] M. Hassaballah, M. A. Kenk, K. Muhammad, and S. Minaee, "Vehicle detection and tracking in adverse weather using a deep learning framework," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 7, pp. 4230–4242, Jul. 2021.

[146] X. Tan, K. Xu, Y. Cao, Y. Zhang, L. Ma, and R. W. H. Lau, "Night-time scene parsing with a large real dataset," 2020, *arXiv:2003.06883*.

[147] H. Ullah *et al.*, "Light-DehazeNet: A novel lightweight CNN architecture for single image dehazing," *IEEE Trans. Image Process.*, vol. 30, pp. 8968–8982, 2021.

[148] T. Hussain *et al.*, "Multiview summarization and activity recognition meet edge computing in IoT environments," *IEEE Internet Things J.*, vol. 8, no. 12, pp. 9634–9644, Jun. 2021.

[149] D. Dai, C. Sakaridis, S. Hecker, and L. Van Gool, "Curriculum model adaptation with synthetic and real data for semantic foggy scene understanding," *Int. J. Comput. Vis.*, vol. 128, no. 5, pp. 1182–1204, 2020.

[150] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic foggy scene understanding with synthetic data," *Int. J. Comput. Vis.*, vol. 126, no. 9, pp. 973–992, 2018.

[151] C. Sakaridis, D. Dai, S. Hecker, and L. Van Gool, "Model adaptation with synthetic and real data for semantic dense foggy scene understanding," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 687–704.

[152] D. Dai and L. V. Gool, "Dark model adaptation: Semantic image segmentation from daytime to nighttime," in *Proc. 21st Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2018, pp. 3819–3824.

[153] C. Sakaridis, D. Dai, and L. Van Gool, "Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3139–3153, Jun. 2022.

[154] O. Zendel, K. Honauer, M. Murschitz, D. Steininger, and G. F. Dominguez, "Wilddash-creating hazard-aware benchmarks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 402–416.

[155] C. Sakaridis, D. Dai, and L. Van Gool, "ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10765–10775.

[156] E. Sadraei *et al.*, "Vehicle–pedestrian interaction: A distributed simulation study," in *Proc. Driving Simul. Conf. Antibes, France*, 2020, pp. 1–8.

[157] J. Zhang, K. Yang, and R. Stiefelhagen, "ISSAFE: Improving semantic segmentation in accidents by fusing event-based data," 2020, *arXiv:2008.08974*.

[158] G. Chen, H. Cao, J. Conradt, H. Tang, F. Rohrbein, and A. Knoll, "Event-based neuromorphic vision for autonomous driving: A paradigm shift for bio-inspired visual sensing and perception," *IEEE Signal Process. Mag.*, vol. 37, no. 4, pp. 34–49, Jul. 2020.

[159] A. Ullah, K. Muhammad, T. Hussain, M. Lee, and S. W. Baik, "Deep LSTM-based sequence learning approaches for action and activity recognition," in *Deep Learning in Computer Vision*. Boca Raton, FL, USA: CRC Press, 2020, pp. 127–150.

[160] A. Hussain, T. Hussain, W. Ullah, and S. W. Baik, "Vision transformer and deep sequence learning for human activity recognition in surveillance videos," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–10, Apr. 2022.

[161] H. Yar, T. Hussain, Z. A. Khan, M. Y. Lee, and S. W. Baik, "Fire detection via effective vision transformers," *J. Korean Inst. Next Gener. Comput.*, vol. 17, no. 5, pp. 21–30, 2021. [Online]. Available: https://www.earticle.net/Article/A402850

[162] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12179–12188.

[163] N. Liu, N. Zhang, K. Wan, L. Shao, and J. Han, "Visual saliency transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4722–4732.

[164] R. Chen *et al.*, "Smoothing matters: Momentum transformer for domain adaptive semantic segmentation," 2022, *arXiv:2203.07988*.

[165] C. Liu, E. Xie, W. Wang, W. Wang, G. Li, and P. Luo, "WegFormer: Transformers for weakly supervised semantic segmentation," 2022, *arXiv:2203.08421*.

[166] R. Liu *et al.*, "TransKD: Transformer knowledge distillation for efficient semantic segmentation," 2022, *arXiv:2202.13393*.

[167] Z. Qin *et al.*, "Pyramid fusion transformer for semantic segmentation," 2022, *arXiv:2201.04019*.

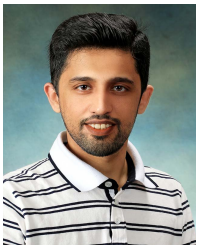
[168] F. Lin, T. Wu, S. Wu, S. Tian, and G. Guo, "Feature selective transformer for semantic image segmentation," 2022, *arXiv:2203.14124*.

[169] T. Hussain, A. Anwar, S. Anwar, L. Petersson, and S. W. Baik, "Pyramidal attention for saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 2878–2888.

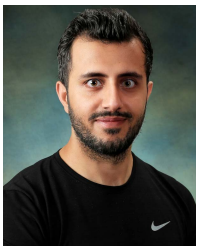
[170] D. Feng *et al.*, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 3, pp. 1341–1360, Mar. 2021.

[171] Y. Zhang, J. Wang, X. Wang, and J. M. Dolan, "Road-segmentation-based curb detection method for self-driving via a 3D-LiDAR sensor," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 12, pp. 3981–3991, Dec. 2018.

Khan Muhammad (Senior Member, IEEE) received the Ph.D. degree in digital contents from Sejong University, Republic of Korea, in February 2019. He was an Assistant Professor with the Department of Software, Sejong University, from March 2019 to February 2022. He is currently the Director of the Visual Analytics for Knowledge Laboratory (VIS2KNOW Laboratory) and an Assistant Professor (Tenure-Track) with the Department of Applied AI, School of Convergence, College of Computing and Informatics, Sungkyunkwan University, Seoul, Republic of Korea. His research interests include intelligent video surveillance, medical image analysis, information security, video summarization, multimedia data analysis, computer vision, the IoT/IoMT, and smart cities. He has registered ten patents and contributed more than 220 papers in peer-reviewed journals and conference proceedings in his research areas. He is an associate editor/an editorial board member for more than 14 journals. He is among the most highly cited researchers in 2021, according to the Web of Science.



Tanveer Hussain (Student Member, IEEE) received the bachelor's degree (Hons.) in computer science from the Islamia College Peshawar, Peshawar, Pakistan, in 2017, and the joint master's and Ph.D. degree from Sejong University, South Korea, in August 2022. He is currently working as a Post-Doctoral Research Fellow in computer vision at the Institute for Transport Studies, University of Leeds, U.K. His major research domains are multimedia data analysis including video summarization, action and activity recognition, saliency detection, scene understanding for autonomous driving, resource-constrained programming, and time series data analysis for power generation and consumption prediction/forecasting. He has filed/published several patents and papers in peer-reviewed journals and conferences in reputed venues, including CVPR, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, IEEE INTERNET OF THINGS JOURNAL, *IEEE Network* Magazine, *Pattern Recognition* (Elsevier), *Neurocomputing*, *Pattern Recognition Letters*, *ACM Computing Surveys*, and *Multimedia Tools and Applications* (Springer). He is a professional reviewer at various reputed journals. He is serving as an Editorial Board Member for the *Journal of Artificial Intelligence and Systems* and a Review Editor for *Frontiers in Artificial Intelligence*. For further activities and implementations of his research, visit: https://github.com/tanveer-hussain.



Hayat Ullah (Student Member, IEEE) received the bachelor's degree in computer science from the Islamia College Peshawar, Peshawar, Pakistan, in 2018, and the master's degree in computer science from Sejong University, Seoul, Republic of Korea, in 2021. He is currently pursing the Ph.D. degree in computer science with Kansas State University, Manhattan, KS, USA. He is also a Research Assistant with the Intelligent Systems, Computer Architecture, Analytics, and Security (ISCAAS) Laboratory, Kansas State University, exclusively working on multi-model human actions modeling and activity recognition. He has published several articles in well-reputed journals, that include IEEE INTERNET OF THINGS JOURNAL and IEEE TRANSACTIONS ON IMAGE PROCESSING. His research interests include image processing, video analytics, deep learning applications in surveillance, applied computer vision, image enhancement, deep reinforcement learning, and image/video quality assessment.

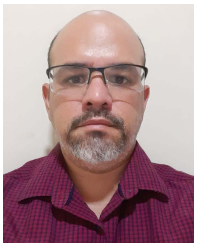
Javier Del Ser (Senior Member, IEEE) received the Ph.D. degree *(cum laude)* in telecommunication engineering from the University of Navarra, Spain, in 2006, and the Ph.D. degree *(summa cum laude)* in computational intelligence from the University of Alcala, Spain, in 2013. He was a Professor and a Researcher at different institutions of the Basque Research Network (including the University of Mondragon, CEIT and Robotiker). He is currently a Research Professor in data analytics and optimization at TECNALIA, Spain, and an Adjunct Professor at the University of the Basque Country (UPV/EHU). His research interests gravitate on the use of descriptive, predictive, and prescriptive algorithms for data mining and optimization in a diverse range of application fields, such as energy, transport, telecommunications, health, industry, and among others. In these fields he has published more than 360 scientific articles, co-supervised 14 Ph.D. thesis, edited six books, coauthored nine patents and participated/led more than 50 research projects. He has also been involved in the organization of various national and international conferences, has chaired three international workshops, and serves as an Associate Editor for a number of indexed journals, including *Information Fusion*, *Swarm and Evolutionary Computation*, and IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS.

Mahdi Rezaei (Member, IEEE) received the Ph.D. degree (Hons.) in computer science from The University of Auckland, New Zealand, in 2014. He is currently an Assistant Professor and a University Academic Fellow (UAF) at the Institute for Transport Studies, University of Leeds. He has 15 years of experience in academia and industry and his primary area of expertise and research interests are in computer vision, machine learning, and deep learning. His research mainly focuses on real-world applications in autonomous driving and smart cars, driver behavior monitoring, road/traffic perception, object detection and tracking, human factors, and safety. He has published more than 50 journal articles and conference papers, a monograph book with Springer, and three best papers. He received the Best Thesis Award 2014 for his Ph.D. degree.

Neeraj Kumar (Senior Member, IEEE) is currently working as a Full Professor with the Department of Computer Science and Engineering, Thapar Institute of Engineering and Technology (Deemed to be University), Patiala, Punjab, India. He is also an Adjunct Professor at King Abdul Aziz University, Jeddah, Saudi Arabia, and Newcastle University, U.K. He has published more than 500 technical research papers in top-cited journals and conferences, which are cited more than 32,700 times with current H-index of 99. His broad research areas are green computing and network management, the IoT, big data analytics, deep learning, and cyber-security. He has also edited/authored ten books with international/national publishers, such as IET, Springer, Elsevier, and CRC. He is serving as an Editor for *ACM Computing Survey*, IEEE TRANSACTIONS ON SUSTAINABLE COMPUTING, IEEE TRANSACTIONS ON NETWORK AND SERVICE MANAGEMENT, *Computer Communication* (Elsevier), and *International Journal of Communication Systems* (Wiley). Also, he has organized various special issues of journals of repute from IEEE, Elsevier, and Springer. Moreover, he won the Best Researcher Award from parent organization every year from last eight consecutive years.

Mohammad Hijji (Member, IEEE) received the Ph.D. degree in computing from Coventry University, U.K., in July 2017. He was the Chairperson of the Computer Science Department, Faculty of Computers, and Information Technology (FCIT), University of Tabuk, Saudi Arabia, from 2020 to 2022, where he is currently the Vice Dean for Development and Quality, FCIT. His research interests include artificial intelligence, cyber security, the Internet of Things (IoT), smart city, energy optimization, disaster, and emergency management.

Paolo Bellavista (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees in computer science engineering from the University of Bologna, Italy. He is currently a Full Professor in distributed and mobile systems with the University of Bologna. His research interests span from pervasive wireless computing to online big data processing under quality constraints and from edge cloud computing to middleware for industry 4.0 applications. He serves on several editorial boards, including IEEE Communications Surveys and Tutorials (an Associate Editor-in-Chief) for *ACM CSUR*, *JNCA* (Elsevier), and *PMC* (Elsevier). He is the Scientific Coordinator of the H2020 BigData Project IoTwins.



Victor Hugo C. de Albuquerque (Senior Member, IEEE) received the bachelor's degree in mechatronics engineering from the Federal Center of Technological Education of Ceará, the M.Sc. degree in teleinformatics engineering from the Federal University of Ceará, and the Ph.D. degree in mechanical engineering from the Federal University of Paraíba. He is currently a Full Professor and a Senior Researcher at the University of Fortaleza, UNIFOR. He leads the Graduate Program in Applied Informatics and Electronics and Health Research Group (CNPq). He mainly researches the IoT, machine/deep learning, pattern recognition, and robotics.