Deep Learning Advances in Computer Vision: A Survey on use of Deep Learning Techniques for Autonomous Vehicles

Dabir Hasan Rizvi
Aberystwyth University
dhr8@aber.ac.uk

***Abstract*—Recent advancements in Deep Learning and Intelligent Transportation System have opened the path for the widespread use of autonomous vehicles. Deep learning has proven to be useful in the design and operation of self-driving cars [1]. Giving rise to new possibilities for intelligent traffic safety, smart roadways, and traveller convenience due to their considerable ability to reduce traffic accidents and human casualties, autonomous vehicles has become a popular area of research [2]. This survey looks at the concept behind autonomous vehicles from a deep learning perspective, as well as contemporary implementations and critical assessments [3]. Through a detailed survey, I hope to bridge the gap between Deep Learning and autonomous vehicles. The overview of autonomous vehicles, overview of modern deep learning technologies, and computer vision are covered in this paper, followed by techniques for object detection and object tracking.**

***Index Terms*—Artificial Intelligence, Autonomous Vehicles, Computer Vision, Deep Learning, Self-Driving Cars**

I. Introduction

Various applications of such technologies have acquired importance because of recent developments in Deep Learning and Computer Vision. Autonomous vehicles are expected to have a significant and fundamental effect on society and the way people travel. Even though acceptance and domestication of technology can be difficult at first, these vehicles will be the first significant integration of personal robots into human civilization in large scale. Automobiles are eventually set to transform into autonomous robots entrusted with human lives, resulting in a diverse socio-economic impact due to rapid advances in Artificial Intelligence technologies. To become a functional reality, these vehicles must have the perception and cognition to deal with high-pressure real-life events, make appropriate decisions, and always take adequate and safe action [3].

Figure 1. depicts the main block diagrams of an autonomous vehicle powered by Artificial Intelligence. The operating choices are taken using either a modular perception-planning-action pipeline (Fig. 1(a)) or an End2End learning approach (Fig. 1(b)), in which sensory data is immediately transferred to control outputs. There are numerous combinations of learning and non-learning-based components that can be used (e.g., a deep learning-based object detector provides input to a classical A-star path planning algorithm). The modular pipeline's components can be built utilising deep learning technologies or traditional non-learning approaches. To ensure that each module is safe, a safety monitor is used [4].

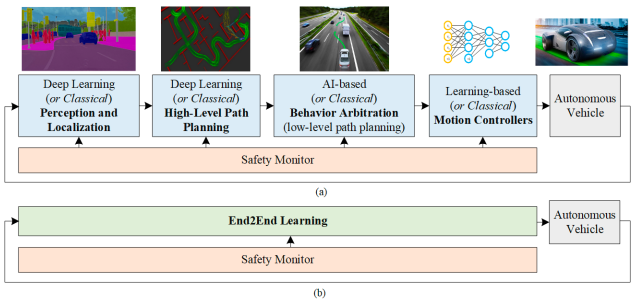

Figure 1: A self-driving car based on Deep Learning. The design can be implemented as a sequential perception planning-action pipeline (a) or as an End2End system (b) [4].

The modular pipeline in Figure 1 is split hierarchically into four components that can be constructed using deep learning and AI technologies or traditional methods. Perception and Localization, High-Level Path Planning, Behaviour Arbitration, or Low-Level Path Planning and Motion Controllers are the components used [4].

Autonomous vehicles can be employed in a variety of situations where having a human driver is inconvenient, unsafe, or impractical. Furthermore, the usage of autonomous vehicles can minimize human errors, lowering the likelihood of road accidents. Several technologies for developing and running autonomous vehicles must improve and develop before this prediction may become a reality. The number of autonomous vehicles on the road will rapidly expand soon, and autonomous vehicles will eventually become the norm in modern traffic. Essentially, autonomous vehicles are intelligent agents that use sensors such as GPS, INS, Lidar, and cameras.

It is also able to interpret and integrate data from many sensors to detect signals and indicators and it is necessary to plan collision-free routes from the initial point to the destination. Autonomous vehicles should be able to detect and manage issues automatically in order to increase safety [1]. In this survey paper, we go through the brief history of evolution of autonomous vehicles and see their benefits and drawbacks to

Submission: 09/05/2022, Contact: dhr8@aber.ac.uk, Aberystwyth University

know about its use in more detail. I have introduced all the major Deep Learning Technologies such as supervised, unsupervised and reinforcement learning and how can it be implemented to autonomous vehicles with the required dataset tools. Computer vision is a major technology used in Autonomous Vehicles and this paper discusses it in detail with focusing on the different methods used for object tracking and object detection. This survey paper discusses the problems, current technological solutions and future work which can improve the use of Computer vision for Deep Learning in Autonomous Vehicles.

II. Overview of Autonomous Vehicles

A. Brief History of Autonomous Vehicles

The concept of self-driving vehicles has been around for about 80 years, with General Motors (GM) Futurama [5] introducing it in the 1939 World's Fair in New York. The emergence of accurate and resilient sensors that continue to shrink in size and cost, along with AI, has been the cornerstone for autonomous driving systems. Self-driving systems based on AI must be able to maneuver effectively in any scenario at any moment. Human-machine interface applications, network-enabled controls, multiple-sensor data fusion, 3D drive scene analysis, and software-defined signal processing are all included in these self-driving systems for transporting goods, payloads, materials, and people [6]. Accurate localization, unobtrusive data collection, fused data-set generation, and uninterrupted high-level communication with other automobiles and neighbouring smart infrastructure are all critical to autonomous navigation accuracy [7]. Self-driving technology is planned to be applied to tractor-trailers, cargo trucks, mining trucks, and buses in the future [3].

Carnegie Mellon University and the Defense Advanced Research Projects Agency (DARPA) have each contributed to the growth of self-driving cars in the recent decade. Google, Tesla Motors, General Motors, Uber, and other automobile businesses anticipate a future with autonomous vehicles [8]. Tesla Motors introduced autopilot technology to its hybrid vehicles, which allowed the cameras and sensors to forecast crashes with up to 76% accuracy, resulting in a collision avoidance rate of more than 90 percent. Numerous infrastructures upgrades include robotic vehicle cruising management systems, automated highway systems, 6G cell-free mobile communications networks with live video processing and no delay are the areas of research that would lead to full-fledged autonomous vehicles creating a greener future [3].

Autonomous vehicles are a highly anticipated technological development with the potential to transform the world. While autonomous cars have gotten a lot of press in the recent decade, driverless transportation has been around for a long period of time [9]. Integration with innovative infrastructure, smart cities, and urban planning with allowances for improved cyber-security, privacy, and insurance are some of the current and larger implications of autonomous vehicles. The extensive application of self-driving technology is exemplified by trains: the SkyTrain in Vancouver, Canada [10], London's Docklands Light Railway (DLR), the Ultra-Pods at London Heathrow Airport [9] Yurikamome in Tokyo, Japan [11] are such examples of automated trains. The mentioned trains operate on closed tracks away from public roadways, avoiding the need to interface with other vehicular traffic or pedestrian [10]. Self-driving cars, on the other hand, are programmed to engage with a wide range of people, potentially resulting in multiple interactions and collisions. The question of whether people will embrace self-driving automobiles as readily as they appear to accept contemporary autonomous transportation is currently being researched [3].

B. Benefits of Autonomous Vehicles

Intelligent Transportation Systems (ITS) use breakthroughs in wireless networking, software-defined networking, and Information and Communication Technology (ICT) to decrease accidents, decrease pollution, alleviate limited mobility, providing newer modes of public transit, and share resources, materials, and space. Such automobiles can be programmed to drive cautiously, avoid blind areas, and adhere to speed regulations [3]. Users may benefit from reduced stress, shorter transit times, productivity improvements, shorter commutes, efficient fuel consumption, and lower carbon emissions. According to research, 1.3 million people die each year because of intoxicated, drugged, distracted, or tired driving. These lives could be saved if autonomous AI systems could eliminate some of the human errors [12].

Self-driving cars would help governments with traffic enforcement, increase road capacity, minimize road fatalities and the frequency of on-road driving-related collisions, and improve speed limit observance. Self-driving vehicles are expected to remove difficulties such as driving while intoxicated, impaired driving, messaging, and other mobile phone use, less braking and acceleration, and highway jams. Fewer accidents are predicted to benefit youngsters and the elderly, allowing people to comfortable and appreciative of self-driving cars. Autonomous electric cars would bring a more environmentally friendly means of transportation, reducing greenhouse and excessive noise while also increasing mobility for the elderly and disabled. Vehicles are parked for long periods of time in today's driving scenario, parking lots can be turned to parks and other green spaces with self-driving automobiles. Self-driving cars would be capable of improving planning and navigation, as well as give the optimum routes to save commute times and costs. While self-driving cars will diminish or perhaps even abolish car ownership, they will enhance shared access while maintaining personalized, effective, and safe transportation [3].

C. Drawbacks of Autonomous Vehicles

Cars are among the most used and easily available means of mobility, and despite advances in technology, driving remains a risky activity. The following are some of the drawbacks of self-driving cars: The primary negative impact of self-driving automobiles would be the loss of jobs in the transportation sector. Self-driving automobiles provide a circumstance in which several lines of source code combined with AI determine a human's life. Even though the role of AI in our society is always changing, an Intelligent system making key judgments must adhere to social norms and respect cultural values to

acquire approval.

It is stated that if a person is in control of an autonomous car or AI system, they will not die or suffer harm if the system fails. The philosophical, ethical, and technological acceptance of self-driving technology is a significant study subject in psychology. Self-driving cars face difficulties at crossings with no traffic signals, faulty traffic lights, uncontrolled intersections, congested crossroads, and areas with pedestrians in close proximity. Self-driving cars are deemed unsuitable for driving in non-mapped locations since they need the global positioning system (GPS) for localization [13]. Because of the car's extensive connectivity, which always keeps it online, it is vulnerable to hacking. Self-driving cars' safety and convenience may jeopardize passengers' privacy because their movements will be monitored and recorded [3].

III. Overview of Deep Learning Technologies

This section introduces deep learning techniques and approaches. Deep learning can be categorized as supervised, unsupervised and reinforcement learning.

A. Supervised Learning

The different types of learning algorithms used in Supervised Learning comprises of Convolution Neural Networks (CNN), Deep Neural Network (DNN), Recurrent Neural Networks (RNN), K-Nearest Neighbour (KNN), etc. The purpose of deep learning during training is to modify the weights of a deep neural network for the model to learn to represent a useful function for its operation. Supervised learning makes use of labeled data in which an expert performs the task at hand. Each observation is approximated by the network during training, and the error is compared to the expert's labelled action. Every data set comprises of an observed action pair that the neural network learns to model [14].

The benefit of using supervised learning includes a faster training convergence and elimination of the requirement to describe the actions to complete the task. The collection and production of data is possible using prior experiences in Supervised Learning and it optimizes performance criteria using previous experiences. The disadvantage of using such algorithm include: Firstly a network which creates predictions about the control action in an offline framework during training, while these predictions have no effect on the states seen during training. When implemented the actions of the network will impact future states, contradiction the assumptions made by majority of the learning algorithms. This causes a distribution shift between training and operation, which causes mistakes from the network due to unfamiliar state distribution observed during operation. Secondly, learning a behavior via demonstration makes the network vulnerable to data set biases. If the goal is to train a generalized model that can drive in all possible environments, the variety of the data set should be ensured for complicated tasks like autonomous driving [15][16].

B. Unsupervised Learning

Unsupervised Machine Learning, analyses and clusters unlabeled datasets using machine learning algorithms. Without the need for human interference, these algorithms uncover hidden patterns or data groupings. It is ideal for exploratory data analysis, cross-selling strategies, consumer segmentation, and picture identification because of its capacity to detect similarities and differences in information. Clustering, Association, and Dimensionality Reduction are the main tasks for unsupervised learning model.

Data mining techniques such as Clustering, groups unlabeled data into groups based on differences and similarities. Clustering algorithms are used to organize raw, unclassified data items into clusters that are represented by information structures or patterns. There are several types of clustering algorithms, including exclusive, overlapping, hierarchical, and probabilistic [17].

Machine learning algorithms have become a popular method to enhance user experience of a product and to test systems for quality assurance. The most common unsupervised learning applications in the real world include news section categorization, computer vision used for visual perception such as object detection, medical imaging such as image detection used in radiology, etc. When compared to manual observation, unsupervised learning gives an exploratory approach to view data, allowing companies to uncover patterns in enormous volumes of data more quickly. While unsupervised learning has numerous advantages, it also has significant drawbacks because it allows machine learning models to run without human interaction [17]. Some of these limitations may include computer complexity due to large amount of training dataset, longer training timeframes, higher chance of inaccurate results, human intervention to validate output variables, and lack of transparency into the foundation on which data was clustered.

C. Reinforcement Learning

Through trial and error, the model can learn to do the task using Reinforcement Learning. It can be modelled as a Markov Decision Process, represented as a tuple (S, A, P, R), where S is the state space. A denotes the action space of feasible actions, P represents the state transition probability model, and R denotes the reward function. At every step, the agent examines a set of states (S), chooses a possible action (A) for that state and the environment adapts according to (P). The agent is then rewarded r_t for a fresh set of states s_{t+1}. The agent's role is to learn a policy $\pi(s_t, a_t)$ which maps the observations to actions in a way that the total rewards are maximized. As a result, an agent can learn its behaviour through its interactions with the environment, and the reward function predicts an estimate of its performance. This method has the advantage of requiring no labelled data sets and allowing reinforcement learning to train a behaviour that generalizes well to varying scenarios.

There are three types of reinforcement learning algorithms: value-based, policy gradient and actor-critic algorithms. Value-based algorithms includes algorithms such as Q-learning that predicts the value function $V(s)$m which indicates the value of predicted reward of the given state. If the state transition dynamics P is given, the policy decided an action that will lead to states which maximizes predicated rewards. In most Reinforcement Learning situations, the environment model is not known. As a result, instead of using the state-action value or quality function *Q (s, a)* that predicts the three values of a particular action in a specific state, the value of state-action and quality function *Q (s, a)* are used. The downside of this method

is that the learned policy's optimality cannot be guaranteed [14]. The disadvantage in Reinforcement Learning is due to its low sample efficiency [18], leading to convergence to an optimal policy can be delayed, necessitating time-intensive simulations or expensive real-world training [19]. Figure 2. represents basic vehicle control system using Reinforcement Learning.

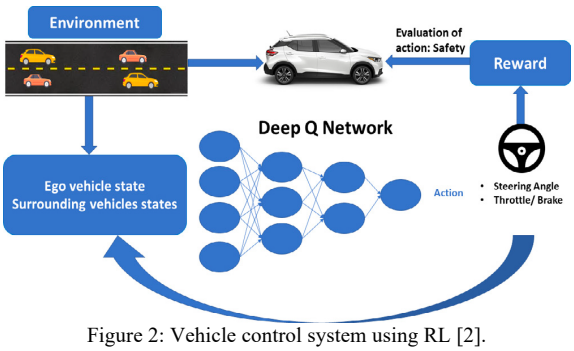

Figure 2: Vehicle control system using RL [2].

Instead of estimating a value function, policy gradient algorithms such as Reinforce parameterizes to maximize predicted rewards it is accomplished by creating a loss function and calculating its gradient with respect to network parameters. The considerable variance in the predicted policy gradient is the fundamental downside of this technique. The network parameters are then modified in the path of the policy gradient during training. Actor-critic algorithms such as A3C are a third-class hybrid method that combine the usage of a parameterized policy function with the usage of value function. The advantage of this technique is that for most success (e.g., reaching a desired location) or failure (e.g., colliding with another item) is simple to define for many tasks. However, because the agent only receives a reward rarely, this can worsen the sample complexity issue in Reinforcement Learning, resulting in slow convergence. In a dense reward function, on the other hand, the agent is rewarded at each time step based on its current state. That suggests, the agent receives a continual learning signal that predicts the utility of the chosen actions in their respective states [14].

D. Datasets Tools for Deep Learning

The rapid advancement in the application of deep learning systems on autonomous vehicles has led to a variety of deep learning data sets for autonomous driving and perception. The KTTI benchmark suite includes different datasets for analyzing stereo vision, optical flow, object detection and tracking, localization and mapping, road detection and semantic segmentation, is the most well-known data set for autonomous driving. Datasets such as Oxford Robotcar, ApolloScape are other useful datasets used for autonomous driving. Aside from publicly available data sets, there are a variety of alternative approaches to developing deep learning in autonomous vehicles [14]. The NVIDIA Drive PX2 is the current leading Artificial Intelligence (AI) platform for autonomous driving [20].

IV. Deep Learning for Computer Vision in Autonomous Vehicles

Computer vision and deep learning are the most important tools for accurately and precisely completing the perception process, including localization and mapping to some extent, with a camera. The camera is a crucial passive sensor for autonomous car's perception of their surroundings. Active sensors include lidar, radar, and sonar to understand their surroundings as we can see in Figure 3. Since cameras are less expensive and more readily accessible, study is conducted on how cameras might be utilized to perceive the world. Monocular and stereo cameras are the two most common camera types [1]. Monocular cameras can retrieve comprehensive information from images in the form of pixel intensities, allowing for the identification of shapes and textural qualities. However, because monocular cameras only have limited depth information, they struggle to estimate an object's position and size. A stereo camera configuration can aid in the resolution of this issue. Time-of-Flight cameras can determine the depth by measuring the time between emitting and receiving infrared pulses [21].

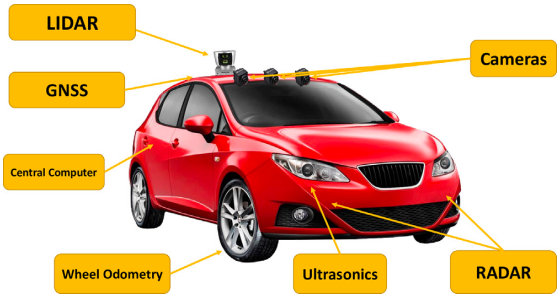

Figure 3: Sensors used in Autonomous Vehicles [2].

Climate and lighting conditions have a variety of effects on camera sensors, and they might cause the sensor to malfunction or produce incorrect photos. The calculation from photos becomes more complex in cold and wet conditions. Lidar provides the most precise measurements at night since depth computation using camera images is inefficient even at night. If an autonomous vehicle's vision sensors fail to recognize objects on the road, it is a severe problem [1]. As a result, the camera is not 100 percent dependable, and lidar is utilized instead, as it is unaffected by light and provides accurate long-distance measurements.

However, lidar is expensive, and that is one of the factors why autonomous vehicles are also costly. Detecting people on the road is a significant challenge for autonomous vehicles. This is a critical responsibility because it involves the safety of persons on the road. Pedestrian detection instrument was proposed that included both an RGB-D stereo vision camera and a thermal camera [22]. Histogram of Oriented Gradient (HOG) and Convolutional Channel Features (CCF) based detection methods were evaluated on a multi-spectral dataset, with CCF outperforming HOG for pedestrian detection. An approach for detecting the number of passengers in a nearby car using monocular vision. In which a series of techniques were used for detection purposes, ranging from detecting a nearby car to detecting passengers inside the windshield, and used a convolutional neural network to accomplish the task was researched [23]. It is critical for an autonomous vehicle to comprehend and recognize the path along which it must travel, as well as to identify and react to objects encountered along that path. Otherwise, it risks causing a dangerous collision if it

deviates from the intended route. When a strong framework is established by autonomous vehicles being able to recognize the domain of drivable path or road, detecting distinct objects on the road becomes easier and more effective [1].

In [24], a Residual Network with Pyramid Pooling (RPP) drivable-road recognition model was proposed, which used a variety of approaches, including convolutional networks, residual learning, and pyramid pooling, to conduct monocular vision-based road detection. When an autonomous vehicle has data on the typical positioning of other road agents in such road intersections, perception and safe positioning become much easier. Road intersections are regarded dangerous, therefore detecting the presence of other cars or road agents is critical. The behaviour of an autonomous vehicle was taught from the extracted visual information about road intersections exiting aerial pictures using a training technique that combined reinforcement learning and computer vision. It was able to investigate the positions of all road agents in the intersections. If a smart city is contemplated, this is an excellent strategy in which road agents learn from aerial images shared amongst themselves and position themselves depending on data obtained from the visual information of aerial photographs. When road intersections are addressed, another issue arises: detecting and understanding traffic signals and acting accordingly. Traffic Light Recognition (TLR) techniques was developed from this concept. The method includes recognizing a traffic light from an image and afterwards deciding based on an evaluation of the light signal's state. The main hurdles in this situation are the sensitive and variable lighting conditions, calculation speed, and dealing with false positives.

In [25], a model for traffic light detection was developed using a High Dynamic Range (HDR) camera for image capture, as HDR cameras have various channels for various exposure values, and then a deep learning neural network with high dynamic imaging, that was possible to identify traffic light in both low exposure/dark frames and high exposure/bright frames. Recently, another remarkable study was conducted on traffic light identification, which used three types of network models centered on the Faster R-CNN and R-FCN models, as well as a six-colour space. They utilized video datasets as input, and because there was fewer yellow light training data in the input, the yellow light was frequently misclassified as red light. As a result, the research is heavily reliant on a huge number of training data sets [1].

V. Object Detection and Tracking

A. Identifying the Issue

The ability to reliably detect objects, as demonstrated in Figure 4, is a prerequisite for autonomous driving. The vehicle utilizes the roadway with many other traffic participants, especially in urban areas, monitoring of other traffic participants or obstructions are required to avoid potential fatal accidents. Because of the wide range of object appearances and occlusions created by other objects or the object of interest directly, detection in metropolitan environments is difficult. Furthermore, physical factors such as cast shadows or reflections, as well as the similarity of objects to one another or the background, can make object detection challenging [26].

Due to their complicated, dynamically variable motion and the broad range of features due to different apparel and dynamic positions, pedestrian recognition is also challenging. Additionally, pedestrian interaction with each other and the environment frequently results in partial occlusions. This issue has been well explored, as demonstrated by advanced driver assistance technologies designed to improve road safety. Pedestrian Protection Systems (PPS) recognize the existence of steady and transient pedestrians in the immediate vicinity of a moving vehicle and alert the driver to high-risk situations. Whereas a driver may still control Pedestrian Protection Systems that miss detections, an autonomous vehicle requires a faultless pedestrian detection system that is reliable in all weather conditions and effective for real-time detection [26].

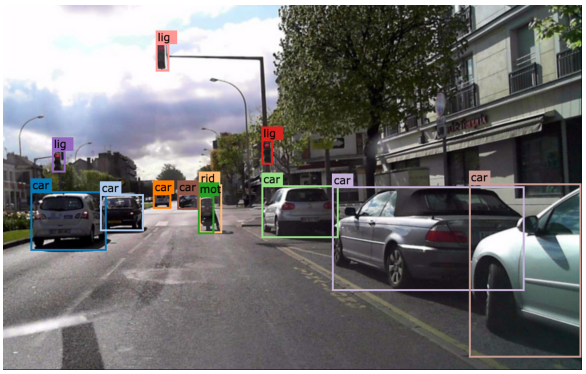

Figure 4: Object Detection, locating all objects in a picture that belong to particular classifications in object detection. Berkeley Deep Drive [27] provided the illustration.

Video cameras are the most affordable and widely used kind of object detection sensor. The visible spectrum (VS) is commonly employed for daytime detections, while the infrared spectrum provides better vision for night-time detections [28]. A range of input modalities have been used to approach the object detection problem. TIR (Thermal Infrared) cameras measure relative temperature, allowing warm objects like pedestrians to be distinguished from cold objects like vegetation or the road. Active sensors, such as laser scanners, that produce signals and observe their reflections, can offer range of information that is useful for detecting and localizing an item in three dimensions. Laser scanners, on the other hand, typically have lower resolution than video cameras. It can be difficult to rely on a single type of sensor alone depending on weather conditions, time of day, or material qualities. Sensor fusion enables for robust integration of this complementary data by combining information from many sensors [26].

B. Classical Pipeline

Pre-processing, Region of Interest Extraction (ROI), object classification, and verification or refinement are typically included in a traditional detection pipeline. Some methods use temporal data in conjunction with a joint detection and tracking system. Typical functions in the pre-processing step include exposure and gain adjustments along with camera calibration and image rectification. A sliding window method, which shifts a window over the image at different scales, can be used to extract regions of interest. Since exhaustive search is costly, numerous techniques for decreasing the search space have been proposed [26].

That is the verification and classification of prospective image regions from sliding windows. Due to the large number of image regions that must be processed, classifying all candidates in an image can be expensive. As a result, a quick choice is needed, that quickly eliminates candidates in the image's background region. Linear Support Vector Machines (SVMs) in combination with Histogram of Orientation (HOG) features have become prominent tools for classification inspired by the work of Dalal and Triggs [29]. Viola et al. [30] use a cascade of simple and efficient classifiers learned using AdaBoost to swiftly eliminate false candidates while focusing more time on promising regions.

C. Part-Based Strategies

Since all possible articulations must be evaluated, learning the appearance of articulated objects is hard. Part-based techniques divide the complicated appearance of non-rigidly moving objects like pedestrians into fewer sections and use these parts to depict articulation, as seen in Figure 5 which makes it flexible and minimizes the amount of training samples required to learn how each element appears. Felzenszwalb et al. [31] developed the Deformable Part Model (DPM), which attempts to simplify the complicated look of objects into simpler elements.

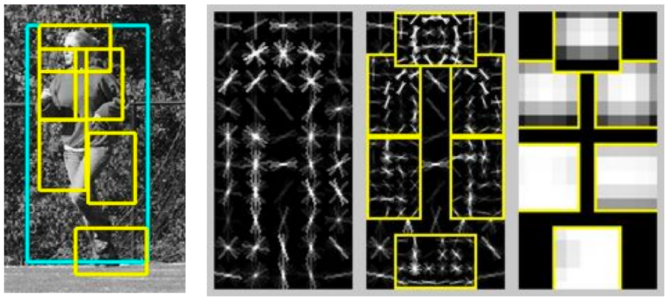

Figure 5: Part-based Approaches. Felzenszwalb et al. [31] proposed the Deformable Part Model (DPM). A coarse global template (middle-left), multiple high-resolution part templates (middle-right), and the location (right) make up the model.

SVM is trained as a classifier using latent structural variables, which reflect the model configuration (part locations) and must be inferred during training. They simulate the appearance of each component using a coarse global template that covers the entire object and higher resolution part templates. A secondary context model is usually learned to place the identified objects in context with the 3D scene. The discussed part-based models are unable to capture relationships between various objects, their components, and the environment, which are required to reason about occlusions [26].

D. Deep Learning for Detection

All prior solutions are based on hand-crafted elements that are hard to develop. Convolutional neural networks are used for object detection tasks with the emergence of deep learning [32], resulting in dramatically improved performance. Figure 6 shows examples of the three most prominent architectures.

Sermanet et al. [33] used convolutional sparse auto-encoders to train the extraction of expressive features in an unsupervised manner, introducing CNNs to the pedestrian detection problem. Furthermore, they use a shallow network with a narrow receptive field that enables for accurate object localization using a sliding window technique. They train an end-to-end supervised classifier while extracting features using a sliding window technique and fine-tuning the auto-encoders together. Deeper networks with bigger receptive fields, on the other hand, make exact localization more difficult since local information is collected in early layers and high-level information is provided in deeper layers. Sermanet et al. [33] developed the first one-stage detector using a deep convolutional variant of the sliding window method. They use a CNN to extract features, then use a sliding window method to apply an AlexNet-based classifier network to the retrieved feature maps. (YOLO) proposes that the topmost feature mappings of a network based on GoogLeNet [34] be used to concurrently learn spatially separated bounding boxes and class probabilities. This enables them to obtain real-time performance, and YOLO9000 to ultimately outperform Region Proposal Networks.

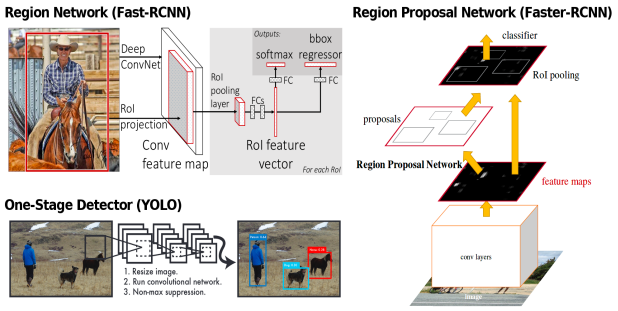

Figure 6: Networks of Object Detection Three typical object detection networks are depicted. Upper left: Fast-RCNN, a region-based network that works on regions. Lower left: YOLO, a one-stage detector that approaches detection as a regression problem. Region proposal network (right) Faster-RCNNs that learn to extract regions faster [26].

By combining feature maps from different sizes and using a fixed set of bounding boxes, Liu et al. [35] increase accuracy and efficiency even more. One-stage detectors, on the other hand, couldn't compete with region proposal algorithms. It provides a dynamically scaled cross-entropy loss to ease this problem and improve training by reducing the contribution of simple samples. The foreground-background class imbalance is one cause for the performance disparity. Anchor bounding boxes are tightly placed over the image, validated, and refined using regression in all preceding one-stage detectors [26].

E. Pedestrian Detection in Real-Time

A quick detection system enables the autonomous system to react quickly in the event of a potential collision with pedestrians. Benenson et al. [36] proposed a quick pedestrian detection system based on improved scale handling and stereo depth extraction. They scale HOG characteristics rather than scaling the photos.

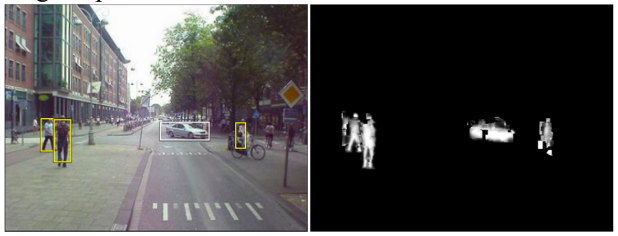

Figure 7: For tracking Object Detections and Segmentations Leibe et al. [37] employed detections (left) and top-down segmentations (right) to learn an object-specific colour model for tracking.

In figure 7, Leibe et al. [37] developed a system for tracking objects and segmentations using object-specific colour model. Furthermore, due to powerful parallelization on the GPU, CNN-based techniques have successfully obtained real-time efficiency. Although Fast R-CNN can just operate at 0.5 Hz, the faster variant, Region Proposal Network Faster-RCNN, was able to operate at 17 Hz. Furthermore, the YOLO9000 can operate at up to 90 Hz with a resolution of 288 × 288 pixels and 40 Hz with a resolution of 544 x 544 pixels [26].

F. Traffic Sign Detection

The most prominent datasets for traffic sign detection are the German Traffic Sign Recognition Benchmark (GTSRB) by Stallkamp et al. [38] and the German Traffic Sign Detection Benchmark (GTSDB) by Houben et al. [39]. Autonomous vehicles require reliable detection and identification of traffic signs. Current CNNs, on the other hand, have already surpassed the constraints of GTSRB and GTSDB, with 100% recall and precision. As a result, Zhu et al. [40] introduced Tsinghua-Tencent 100K, a new traffic sign identification benchmark, to the community, posing additional problems. SVMs, pattern matching approaches, voting schemes such as radial symmetric detectors, and integral channel features have all been investigated for traffic sign detection.

VI. Conclusion and Future Directions

This survey paper provides an overview of recent Deep Learning breakthroughs in computer vision for autonomous vehicles. The evolution of autonomous vehicles in recent history and its advantages and disadvantages is discussed in this survey paper. Different types of Deep Learning techniques used currently and the use of computer vision in autonomous vehicles has led to great developments in the field of Artificial Intelligence. The survey paper covers different techniques used for Object Detection and Object Tracking for autonomous vehicles. Even though autonomous vehicles have a long history, predicting when they will reach the global market remains difficult. Historically, the difficulties associated with approaching or exceeding human-level performance on this task has been neglected. Most self-driving systems depend on precise HD maps for localization and recognition of static infrastructure, which are difficult to generate and keep up to date. The great accuracy that must be achieved, the robustness needed for safe self-driving, and bad weather conditions are all obstacles. Pedestrians are another hurdle for self-driving vehicles, as their action is often unpredictable, and interacting with them is crucial for making driving judgments. Before self-driving cars may be used globally on public highways, various ethical and legal issues must be resolved [26].

With the introduction of Artificial Intelligence based autonomous vehicles on public roadways, several challenges were faced. Given the existing framework and interpretability of neural networks, demonstrating the safety and reliability of autonomous vehicles is a huge task. Deep learning algorithms also rely on big training databases and demand a lot of processing power [4]. Due to CNN's outstanding ability to operate as feature extractors, an examination of existing deep learning architectures, frameworks, and models found that CNN and a mix of RNN is currently the most employed technique for object detection.

Future works to improve self-driving car include:

- To collect reliable data in dangerous weather situations such as rain, hail, and snow, and to research autonomous vehicle's navigation without human interference.
- Techniques for Single-Shot Detectors can be adapted to operate with videos. A dataset with vehicle videos captured at 40 frames per second (fps) was recently provided and can be used to test current cloud-based Deep Learning in real-time [41].
- Understanding and segmenting driving scenes can be combined with information transmission through time to comprehend both space and time.
- RNNs and other deep learning architectures can be used to perform automatic picture labelling, localization, and detection.
- Deep Learning algorithms are unable to transfer representations to unrelated domains. Transfer learning between domains is one research route that could lead to innovative ways to analyze scenes and real-time object recognition [3].

References

[1] J. Ren, H. Gaber, and S. S. A. Jabar, "Applying deep learning to autonomous vehicles: A survey," in *Proc. 4th Int. Conf. Artificial Intellifence. Big Data (ICAIBD)*, May 2021, pp. 247–252.

[2] B. B. Elallid, N. Benamar, A. S. Hafid, T. Rachidi, AND N. Mrani, "A Comprehensive Survey on the Application of Deep and Reinforcement Learning Approaches in Autonomous Driving", *Journal of King Saud University-Computer and Information Sciences*, 2022.

[3] A. Gupta, A. Anpalagan, L. Guan, and A. S. Khwaja, "Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues," Array, vol. 10, p. 100057, 2021.

[4] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, "A survey of deep learning techniques for autonomous driving," Journal of Field Robotics, vol. 37, no. 3, pp. 362–386, 2020.

[5] H. Wasson, "The Other Small Screen: Moving Images at New York's World Fair, 1939," Canadian Journal of Film Studies, vol. 21, no. 1, pp. 81–103, 2012.

[6] E. Cho and Y. Jung, "Consumers' understanding of autonomous driving," Information Technology & People, 2018.

[7] E. Laes, L. Gorissen, and F. Nevens, "A comparison of energy transition governance in Germany, the Netherlands and the United Kingdom," Sustainability, vol. 6, no. 3, pp. 1129–1152, 2014.

[8] M. Ryan, "The future of transportation: ethical, legal, social and economic impacts of self-driving vehicles in the year 2025," Science and engineering ethics, vol. 26, no. 3, pp. 1185–1208, 2020.

[9] S. Nordhoff, J. De Winter, M. Kyriakidis, B. Van Arem, and R. Happee, "Acceptance of driverless vehicles: Results from a large cross-national questionnaire study," Journal of Advanced Transportation, vol. 2018, 2018.

[10] R. D. Robertson, S. R. Meister, W. G. Vanlaar, and M. M. Hing, "Automated vehicles and behavioural adaptation in Canada," Transportation Research Part A: Policy and Practice, vol. 104, pp. 50–57, 2017.

[11] S. Deb, L. Strawderman, D. W. Carruth, J. DuBien, B. Smith, and T. M. Garrison, "Development and validation of a questionnaire to assess pedestrian receptivity toward fully autonomous vehicles," Transportation research part C: emerging technologies, vol. 84, pp. 178–195, 2017.

[12] J. C. De Winter, R. Happee, M. H. Martens, and N. A. Stanton, "Effects of adaptive cruise control and highly automated driving on workload and situation awareness: A review of the empirical

evidence," Transportation research part F: traffic psychology and behaviour, vol. 27, pp. 196–217, 2014.
[13] R. W. Wolcott and R. M. Eustice, "Robust LIDAR localization using multiresolution Gaussian mixture maps for autonomous driving," The International Journal of Robotics Research, vol. 36, no. 3, pp. 292–319, 2017.
[14] S. Kuutti, R. Bowden, Y. Jin, P. Barber, and S. Fallah, "A survey of deep learning applications to autonomous vehicle control," IEEE Transactions on Intelligent Transportation Systems, vol. 22, no. 2, pp. 712–733, 2020.
[15] A. Torralba, A. A. Efros et al., "Unbiased look at dataset bias." In CVPR, vol. 1, no. 2. Citeseer, 2011, p. 7.
[16] A. Gupta, A. Murali, D. P. Gandhi, and L. Pinto, "Robot learning in homes: Improving generalization and reducing dataset bias," in Advances in Neural Information Processing Systems, 2018, pp. 9094–9104.
[17] IBM Cloud Education (Ed.). "What is unsupervised learning?" IBM. Retrieved from: https://www.ibm.com/cloud/learn/unsupervised-learning,2021
[18] Z. Wang, V. Bapst, N. Heess, V. Mnih, R. Munos, K. Kavukcuoglu, and N. de Freitas, "Sample efficient actor-critic with experience replay," arXiv preprint arXiv:1611.01224, 2016.
[19] R. S. Sutton and A. G. Barto, :Reinforcement Learning: An Introduction," Cambridge, MA: MIT Press, 1998, vol. 9.
[20] NVIDIA Corporation, "Autonomous car development platform from NVIDIA DRIVE PX2," 2018. [Online]. Available: https: //www.nvidia.com/en-us/self-driving-cars/drive-platform/
[21] 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, Oct. 2019.
[22] Z. Chen, and X. Huang, "Pedestrian Detection for Autonomous Vehicle Using Multi-Spectral Cameras," IEEE Transactions on Intelligent Vehicles, vol. 4, No. 2, Jun. 2019.
[23] A. Amanatiadis, E. Karakasis, L. Bampis, S. Ploumpis, and A. Gasteratos, "ViPED: On-road vehicle passenger detection for autonomous vehicles," Robotics and Autonomous Systems, Dec. 2018.
[24] X. Liu, and Z. Deng, "Segmentation of Drivable Road Using Deep Fully Convolutional Residual Network with Pyramid Pooling," Cognitive Computation, Nov. 2017.
[25] J. Wang, and L. Zhou "Traffic Light Recognition With High Dynamic Range Imaging and Deep Learning," IEEE Trans. Intell. Transp. Syst, vol. 20, no. 4, Apr. 2019.
[26] J. Janai, F. Güney, A. Behl, A. Geiger, and others, "Computer vision for autonomous vehicles: Problems, datasets and state of the art," Foundations and Trends® in Computer Graphics and Vision, vol. 12, no. 1–3, pp. 1–308, 2020.
[27] H. Xu, Y. Gao, F. Yu, and T. Darrell, "End-to-end learning of driving models from large-scale video datasets," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2174–2182.
[28] F. Suard, A. Rakotomamonjy, A. Bensrhair, and A. Broggi, "Pedestrian detection using infrared images and histograms of oriented gradients," in 2006 IEEE Intelligent Vehicles Symposium, 2006, pp. 206–212.
[29] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), 2005, vol. 1, pp. 886–893.
[30] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," International Journal of Computer Vision, vol. 63, no. 2, pp. 153–161, 2005.
[31] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in 2008 IEEE conference on computer vision and pattern recognition, 2008, pp. 1–8.
[32] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in neural information processing systems, vol. 25, 2012.
[33] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun, "Pedestrian detection with unsupervised multi-stage feature learning," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pp. 3626–3633.
[34] C. Szegedy et al., "Going deeper with convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.
[35] W. Liu et al., "Ssd: Single shot multibox detector," in European conference on computer vision, 2016, pp. 21–37.
[36] R. Benenson, M. Mathias, R. Timofte, and L. Van Gool, "Pedestrian detection at 100 frames per second," in 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2903–2910.s
[37] B. Leibe, K. Schindler, N. Cornelis, and L. Van Gool, "Coupled object detection and tracking from static cameras and moving vehicles," IEEE transactions on pattern analysis and machine intelligence, vol. 30, no. 10, pp. 1683–1698, 2008.
[38] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "The German traffic sign recognition benchmark: a multi-class classification competition," in The 2011 international joint conference on neural networks, 2011, pp. 1453–1460.
[39] S. Houben, J. Stallkamp, J. Salmen, M. Schlipsing, and C. Igel, "Detection of traffic signs in real-world images: The German Traffic Sign Detection Benchmark," in The 2013 international joint conference on neural networks (IJCNN), 2013, pp. 1–8.
[40] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu, "Traffic-sign detection and classification in the wild," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2110–2118.
[41] M. U. Yaseen, A. Anjum, M. Farid, and N. Antonopoulos, "Cloud-based video analytics using convolutional neural networks," Software: Practice and Experience, vol. 49, no. 4, pp. 565–583, 2019.