ISSN No:-2456-2165

Modelling Autonomous Driving and Obstacle Avoidance Using Multi-Modal Fusion Transformer Framework

Abhinav Singh
Department of CSE Dayananda Sagar College of Engineering, Bangalore

Nishant Prakash
Department of CSE Dayananda Sagar College of Engineering, Bangalore

Kanishk Bhaskar
Department of CSE Dayananda Sagar College of Engineering, Bangalore

Krishnan Rangarajan, Professor
Department of CSE Dayananda Sagar College of Engineering, Bangalore

Aditya Raj
Department of CSE Dayananda Sagar College of Engineering, Bnagalore

Abstract:- The papers that we are surveying have many methods that have been presented with different solutions for autonomous driving. One of the few novel representations helps in proving the reasoning for imitation learning in a certain scene where the cameras are used to highlight a certain location which coordinates to waypoints and semantics. In this method the camera follows the car and will show the waypoints at a certain distance ahead of the car at all times while the car is moving. The papers have used attention fields to compress two-dimensional images with features which are best suited for cognitive processing on a discrete aspect of information or in other words obstacles that may appear in front of the car. Therefore, the other model being a Multi-Modal Fusion Transformer is used to combine two separate datasets such as image data and topography data from cameras and distance sensors respectively using attention mechanism. This helps in integrating image data and the topography data that is being received through the camera and distance sensors. The distance sensor maps the surface of all the surroundings where the car is being driven.

Keywords:- *End-to-End Autonomous Driving, Transformer, 2D Imaging, Self-Attention Model, Imitation Learning.*

I. INTRODUCTION

Through this survey we have noticed the workings of high-performance, Aautonomous driving models using self-attention models. One of the models continuously maps obstacles and locations in a certain view where it will continuously follow the car and determine the waypoints which are ahead of the car and report the coordinates and the position of the obstacle that is front of the vehicle. The second model, a Multi-Modal Fusion Transformer, helps combine two separate datasets that are gathered for the model training such as image and topography representations through attention mechanisms. Through these research papers, we saw different ways that can be used to counter problems such as BEV semantic prediction and vehicle trajectory planning from the inputs that we will be getting through cameras and lidar sensor. Since the model with BEV sematic view is flexible in input modalities and output supervisions, it is combined with reinforcement learning for adjusting weights to counter any errors that may hinder the model's performance. It also demonstrates the imitation and certain learning policies based on these novels that provides us with the solutions on how to deal with complex scenarios such as Traffic Control for multiple vehicles, multiple vehicle movements at uncontrolled intersections from various directions, pothole detection and speed breaker detection. We have also noticed development of counter measures to certain problems mentioned before so that the model does not under perform in these scenarios. We have learned about the measures to check and validate the accuracy of these models which were used in many different environment settings with multiple complex scenarios using a simulator to drive the framework.

Through the research paper we discovered that 2 sensors cameras which projects different waypoints and LiDAR or other Distance Sensors are being used to gather the data. With the camera, image data can be obtained and with LiDAR and distance sensor the topographic data can be obtained. With two significant data sources they are used as the input through the transformers and to the convolutional neural network using the models. After pre-processing the dataset and training the models, network computer commands are obtained and the performance is measured through a metric score.

Output representation – Prediction of the vehicle movement and trajectory in a semantic space representation in accordance with the current coordinates of the vehicle.

II. LITERATURE SURVEY

- *Implicit Scene Representations*

Using neural implicit scene representation for classifying the geometry or the topography or the surface of the area where the car is moving. These methods depicts that the surfaces or the area of the topography is considered as the limit for the neural classifier. They have been applied for representing the different object dynamics and general lighting properties. In the papers, we have surveyed that there are two important detection methods to be used which are object detection and lane detection. In the object detection method, in which we obtain high-resolution scene representations while remaining in a stable condition. Now for it we have seen that using of cameras will be beneficial. For which there will be one camera being used as a bird eye view for detecting speed breakers and potholes. While we will have other 3 cameras on the left, right and the centre of the model for giving us the constant memory footprint of the image data. The other method we have examined and that can be used is lane detection. For that the vehicle will be constantly be checked whether the model have followed the lane and changed its position only if it is required. While surveying it was noticed that one of the models is provided by the same properties after the examination and will check the error between the predicted outputs and the expected output and minimize the errors using neural approximation for creating a better learning environment for the model for autonomous driving tasks.

- *End-To-End Autonomous Driving*

Certain learning based methods used Autonomous Driving are still under research. It has been used for driving with an advanced in an important way and is currently used in multiple complex scenarios in real world. Those approaches we examined were the waypoints predictions and the others approaches were to directly able to the predict vehicular control. While other methods that we had examined were the method of learning-based driving method which includes features such as affordances and the Reinforcement Learning. This method could surely benefit from an encoder which projects image data in a specified semantic view. So, in this work, we examined that we can apply this encoder for image gathering and improve Imitation Learning based autonomous driving.

- *BEV Semantics For Driving*

BEV Semantics or also known as bird’s eye view uses a top-down view of a street as it gives a close to accurate 3D perception of a route/street which are essential for autonomous driving related task. It generally gives a three-dimensional layout of the area where the vehicle is present. This gives a rough projection of the three-dimensional space which is later used to detect obstacles in front of the car during complex scenarios. For example, (LBC student-teacher model) a teacher uses bird’s eye view representation to learn driving, the generated input is the used to supervise a student aiming to perform the same task but only using images. By doing so the system achieves a performance boost which deems it slightly better than the previous CARLA versions, showcasing the benefits of BEV representations.

- *Imitation Learning*

Imitation Learning learns a policy which is basically a network computed command. It essentially imitates the behaviour of an expert or a human behaviour. In imitation learning, a policy basically means mapping from various inputs such as BEV or other similar representations and topography of the area which is projected in a layout format to waypoints that are provided to transformer as a common input and send to Multi-layer perceptron which acts as a separate controller. One of the approaches of the imitation learning is a supervised learning method called Behavior cloning. Through this we are deriving Dataset ‘D’ of size ‘Z’.

$$D = \{(X^i, W^i)\}_{i=1}^{Z}$$

which consists of high-dimensional observations of the environment ‘X’ and the corresponding expert trajectory, defined by a set of 2D waypoints ‘W’ in BEV space, i.e.,

$$W = \{w_t = \left(x_{t,}\, y_t\right)_{t=1}^{T}$$

The specific semantic view representation that is being sought uses the cameras in the vehicle to pin-point the coordinates of the vehicle. The policy, which are basically network computed commands are trained through supervised learning method using data ‘D’ with the loss function L.

In this observation, we see the use of cameras for gathering image data and the use of distance sensors for gathering topography data for a single time frame ‘t’. The use of single time frame as input is more beneficial on imitation learning for autonomous driving because surveying from other research papers and reports have concluded that data collected from previous observation or reports may not be helpful if the model is to gain in performance. We have seen the use of ‘L’ which is the loss function and states the distance between the predicted movement of the vehicle and the expected movement of the vehicle ‘W’, as the loss function.

- *Multi-Modal Fusion Transformer*

The main idea of a multi-modal fusion transformer is the self-attention mechanism which allows the model to select the desired features automatically for the autonomous driving framework. The idea is to generate a layout format of the topography or the area in which the car is being driven using LiDAR. The LiDAR uses a near-infrared laser to map the topography it is scanning. The transformer takes input of both the images data collected using cameras and the topography data using LiDAR or other Distance sensors. Such inputs are considered as tokens to get a feature vector.

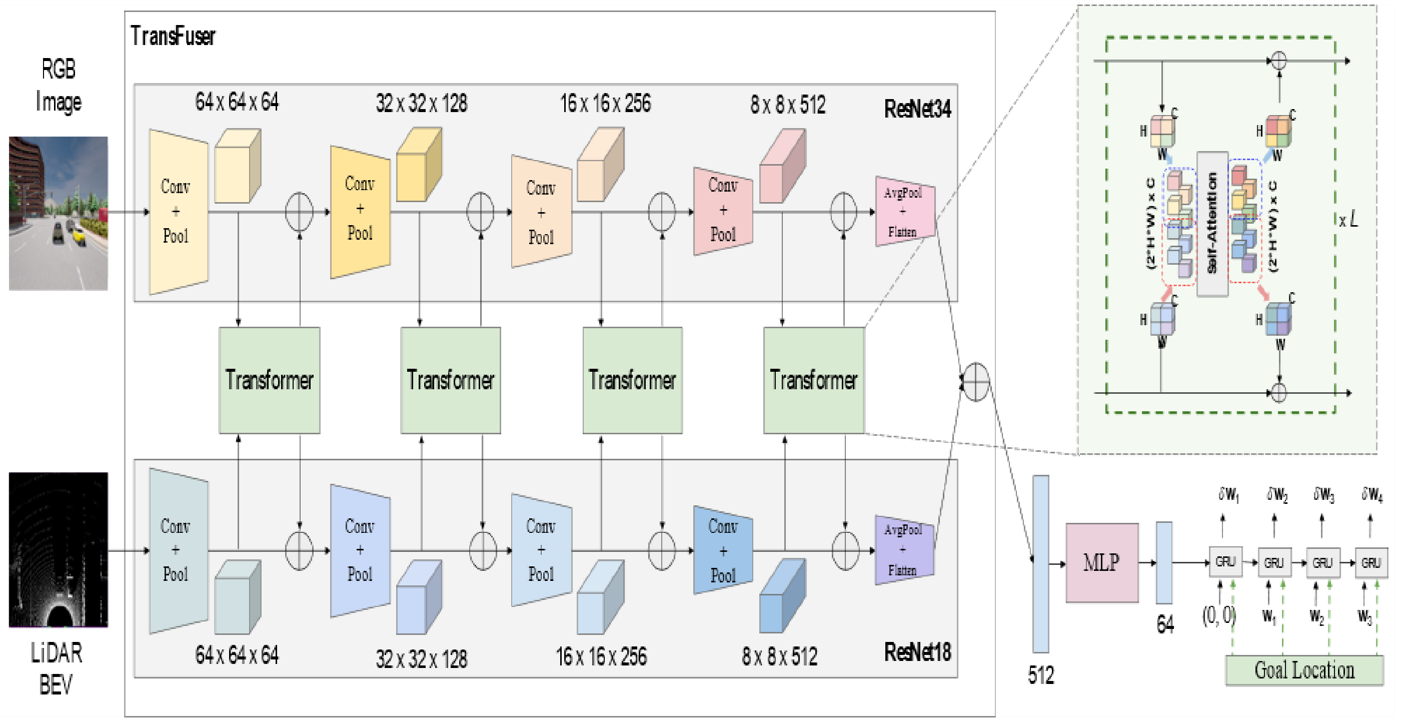

Fig 1 Architecture

- *Waypoint Prediction Network*

Waypoint prediction network is constructed in a way where we are able to determine various obstacles which are in front of the vehicle. The waypoints are generally coordinates of the obstacles that appear in front of the vehicle and is continuously mapped in real time. The data gathered from the cameras and the distance sensors are combined and passed to the fusion transformer which in turn transfers the input in the form of ResNet to the multi-layer perceptron. After passing it to the multi-layer perceptron, the model is trained and the network computed commands are generated which performs the autonomous driving tasks.

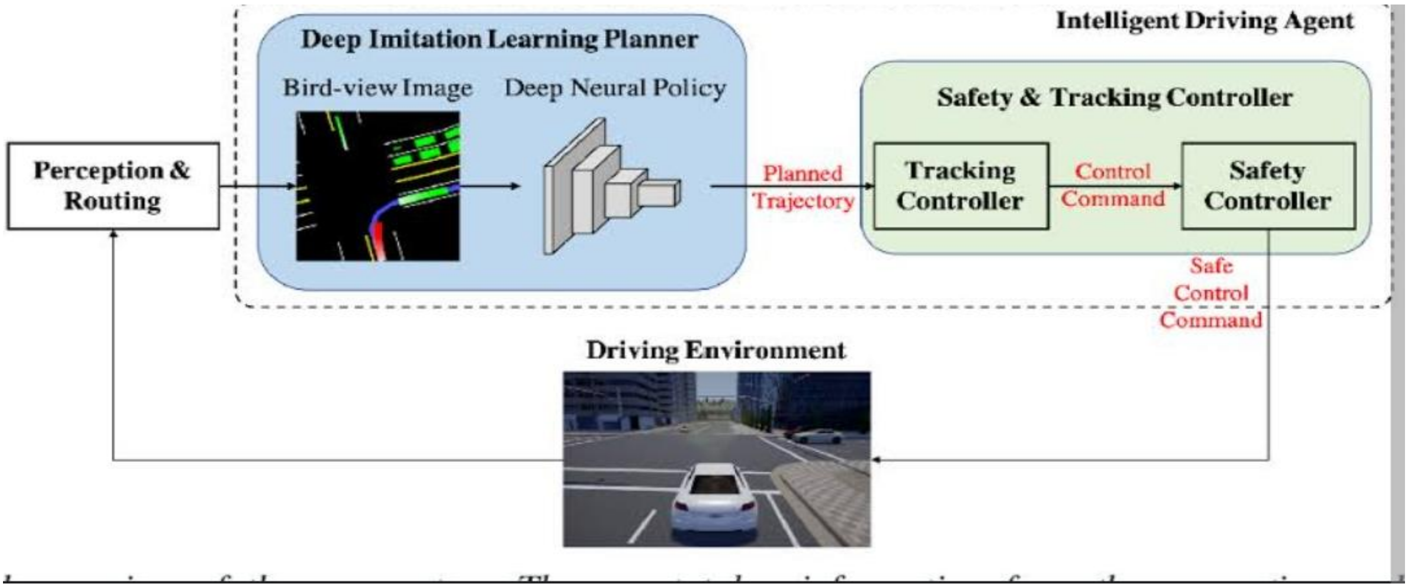

Fig 2 This figure represents an overview of one of the systems of the autonomous driving framework and it's working using Deep imitation learning and object tracking. Through this we are deriving certain policies for driving tasks.

- *Loss Function*

At last, we can train the network using an 'L' loss function which will actually show the difference between the predicted output and the expected output after model testing, from the data which was being used and the current coordinate of the vehicle. So now we can use the variable 'w_t' to represent the prediction output with the time frame 't'. And then we can also use the variable 'w^gt_t' to represent the expected output for the time frame 't'. Then the loss function that will be calculated using the summation of both values and can be represented in the mathematical representation which is given as follows:

$$L = \sum_{t=1}^{T} \|w_t - w_t^{gt}\|_1$$

- *Task*

We have seen the use of CARLA(version 0.9.10) to run the navigation simulation along predefined routes. A sequence of sparse GPS locations defines the routes on which the simulations will run. An agent needs to complete the route while registering dynamic background agents(eg. pedestrians, cyclists, vehicles) and traffic rules. Each of the route may contain several scenarios(eg. Multiple vehicular movements at uncontrolled intersections, vehicles jumping red lights or breaking traffic rules by not detecting the traffic signs, sudden appearance of pedestrians or other obstacles while the car is moving).

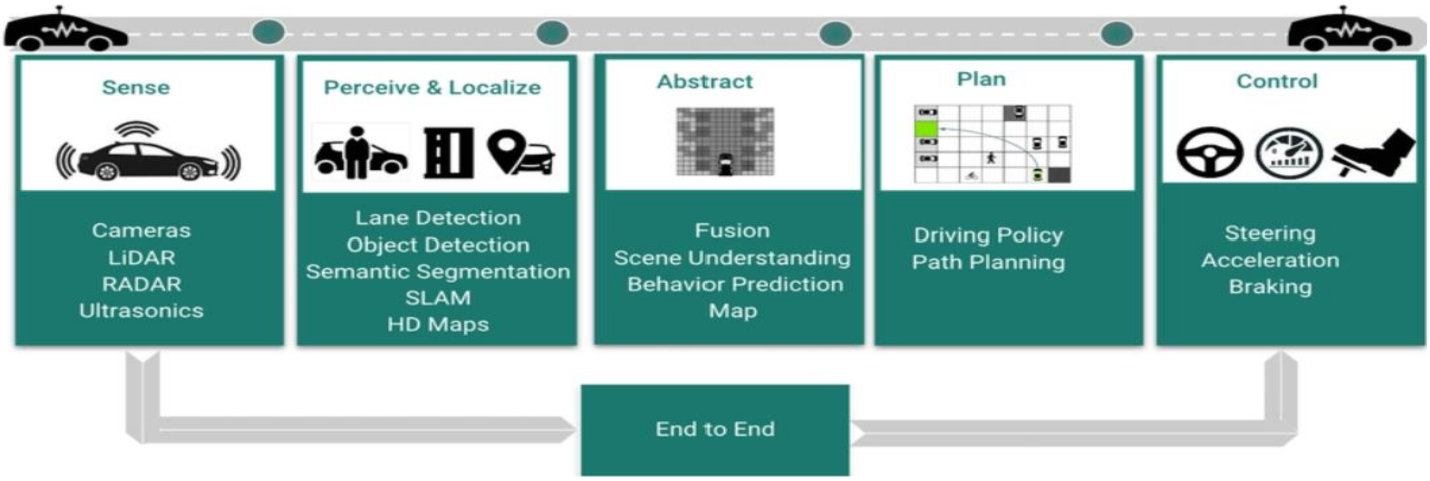

Fig 3 Modules in an autonomous driving pipeline.

- *Metrics*

Through various research papers, we have deduced some of the metrics that can be used to judge the performance of the autonomous driving framework such as Route completion score, Infraction Score and Driving Score. If we consider one particular route, the route completion can be calculated by checking the percentage of the total distance completed by the vehicle from the start to the end of a particular route. During this, the framework will also be judged based on how many times did it break the traffic rules and at what moment was it not able to deduce a solution to the problem. The Infraction score can be calculated by how many times has the vehicle deviated from the lane, traffic sign violation and various other violations caused during the driving. However, the driving score can be calculated as the route completion metric is weighted by the infraction score for that particular route. Similarly, the metrics can be calculated for all other routes with different environments. After calculating the mean of all the metrics for all the environments, a report can be generated to formulate the standard deviation and mean after conducting internal evaluations on the models individually multiple times.

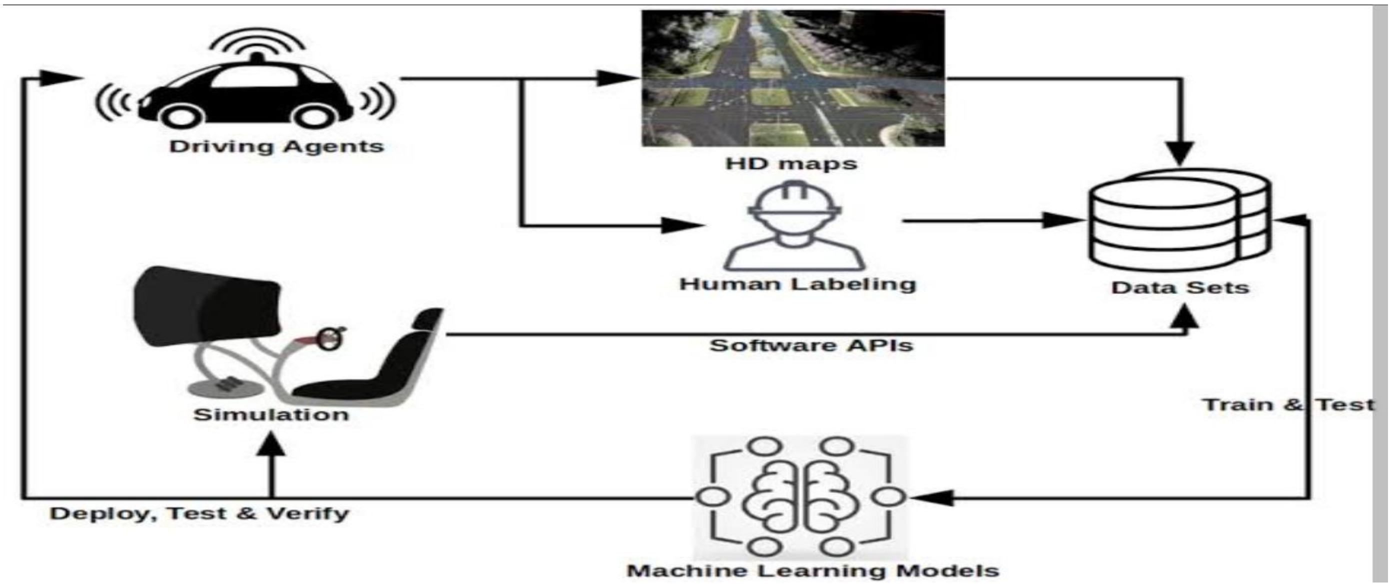

Fig 4 This figure shows the pipeline or the framework that is being used to create an autonomous driving model.

III. COMPARISION TABLE

The comparison table has been made to demonstrate and compare the working and the methodologies of different research and journal papers that have been used in this survey. Through this we were able to analyze and conclude the best possible approach for end-to-end autonomous driving. The papers have briefly discussed and shown their approach to solve multiple complex scenarios and how best to enhance the model's performance.

Table 1 Comparision

Author, year	Title	Remarks
J. Janai, 2017	Computer vision for autonomous vehicles: Problems, datasets and state of the art	In this paper, we analyzed the performance of the autonomous driving model using simple computer vision framework on several benchmarking datasets, including KITTI, MOT, and Cityscapes. We are checking the overview of datasets for the model by making perception training as the primary focus of the dataset. Then we noted updates on the object detection, reconstruction, and scene understanding techniques to create algorithms to counter the problems. To solve these problems, we see that the paper uses an interactive online tool which form a mental image of the surveyed papers which also consists of an interactive graph and some additional information to be used in it.
A. M. López, 2017	Computer Vision in Vehicle Technology: Land, Sea & Air. Hoboken	In this paper we have examined that the author uses computer vision framework for driver assistance and robotic navigation. Computer Vision is being used in the field of autonomous driving more than ever. It is heavily used in drones and cars for moving the vehicle and projecting the trajectory. It is also constantly being used in the area of research and development for multiple applications and also for major developments in technology.
H. Xu, 2017	End-to-end learning of driving models from large-scale video datasets	In this paper we analyzed the learning of motion of the vehicle using data gathered from cameras. The data is based on the video format at certain fps. This helps in training the end-to-end framework of the autonomous driving model. The model uses a novel FCN-LSTM framework, that can be extracted from vehicle motion data and use these data for scenic segmentation which is essential for improved performance. A large-scale dataset is essential to train the driving behavior and will also help in predicting driver action on multiple environments and conditions.
B. Yang, 2018	Real-time 3D object detection from point clouds	In this paper we have examined the recent approach of 3D object detection in real time which is PIXOR. This is the object detection which make the full use of 2D BEV representation in a maximum productivity with minimum wasted efforts. The two datasets on which the PIXOR runs are KIITI benchmark and a real time 3D vehicle detection dataset. The datasets that is being presented shows that the proposed detection architecture surpasses other methods in terms of Average Precision (AP) which deems it slightly better in terms of model accuracy, while still running on more than 28 FPS.

Yilun Chen, 2019	Attention-based hierarchical deep reinforcement learning for lane change behaviors in autonomous driving	In this paper we have examined that the author designs a Deep Reinforcement Learning algorithm to learn lane changing behaviors of a vehicle in closed or different traffic environment. The author shows a method for learning the different driving policies into a single model. They also Help the above method by using the image data for improving the accuracy for the model to work. They also work on to apply spatial attention to the Deep reinforcement Learning architecture. This helps the vehicle adapt and look around it's area of trajectory for other vehicles and determines the lane changing behavior based on the current traffic condition. The experiments are conducted in the TORCS simulator and the results have been proven to be better than other deep reinforcement learning frameworks that perform similar lane changing feature.
E. Santana, 2017	Learning a driving simulator	In this paper, we have examined that the authors illustrate one of CommaAI's research approaches for driving simulation. One where they learn to simulate. Here the authors try to investigate and work on the variational autoencoders with classical and learned cost functions using the generative adversarial networks for embedding road frames. Afterwards, a transition model is learned in the embedded space using action-conditioned Recurrent Neural Networks (RNNs). It then shows that the approach can be kept predicting realistic looking video for several frames despite the transition model being optimized without having any cost function in the pixel space provided to us.
Z. Yang, 2018	End-to-end multimodal multi-task vehicle control for self-driving cars with visual perceptions	In this paper, we have examined that the authors propose a multi-task learning framework to predict the steering angle and speed control simultaneously in a simulation environment. The first network helps in predicting the steering angle with respect to the speed of the vehicle. Moreover, the authors then propose a multi-modal network to predict speed of the vehicle and steering angles by taking previous speeds and visual recordings as input for the model. The following experiments are conducted on the public Udacity dataset and a newly collected SAIC dataset. More on this paper we also check the failure of the vehicle to guess the accurate steering angle input data which is then used to synthesize new solutions to counter incorrect steering input.

IV. CONCLUSION

Through this survey, we have been able to learn about certain IL[01] policies. We have been able to comprehend the scope of these policies and their uses in end-to-end autonomous driving. Few methods that we have been able to survey are Waypoint prediction that will use a camera and constantly follow the car to generate image data of the road ahead of the car for a certain distance in a different scene representation. The other method that is being used is a multi-modal fusion transformer which has as self-attention mechanism that uses the input modalities of both cameras for the image data and distance sensors for the topography data and uses it as a common input for the transformers being used for the model. One other method that is being used is imitation learning where a policy is generated through the model after network computed command is generated. This is used by mapping the input modalities to the waypoints from a different controller to generate output. All these are just a few methods that have been devised to perform end-to-end autonomous driving at different places such as uncontrolled intersections with multiple vehicles, Pothole detection. It also detects vehicles, objects and potholes if there is blind spot in the area or if the objects are not visible to the driver.

For data generations, we have seen the use of an expert agent on multiple publicly available towns in CARLA. The paper evaluates the models with different approaches, which consists of various other unknown routes and environment conditions. It also conducts an internal evaluation with a combination of multiple and various environments. Each route has a unique environmental condition, which are a combination of various weather conditions(eg. clear, cloudy, wet, mid-rain, wet-cloudy etc.) with one of multiple daylight conditions(e.g. night, twilight, dawn, morning, noon, sunset). Additional features can also be added to test the model's accuracy over other environments and make navigating through these environments more challenging so that the scores of model's can be analyzed and be compared to previous results.

REFERENCES

[1]. Yi Xiao, Felipe Codevilla, Akhil Gurram, Onay Urfalioglu, and Antonio M. López, Member, IEEE: Multimodal End-to-End Autonomous Driving.

[2]. Eloi Zablocki, Hedi Ben-Younes, Patrick Perez, Matthieu Cord: Explainability of deep vision-based autonomous driving systems: Review and challenges.

[3]. Aditya Prakash, Kashyap Chitta, and Andreas Geiger: Multi-Modal Fusion Transformer for End-to-End Autonomous Driving

[4]. Jianyu Chen, Bodi Yuan and Masayoshi Tomizuka: Deep Imitation Learning For Autonomous Driving In Generic Urban Scenarios With Enhanced Safety.

[5]. Chen Sun, Jean M. Uwabeza Vianney, and Dongpu Cao: Affordance Learning In Direct perception for autonomous Driving.

[6]. Adithya Ranga, Filippo Giruzzi, Jagdish Bhanushali1, Emilie Wirbel, Patrick P´erez, Tuan-Hung Vu, Xavier Perotton: Multi-Task Learning Model For Intent Prediction Of Vulnerable Road Users.

[7]. A. M. López, A. Imiya, T. Pajdla, and J. M. Álvarez, Computer Vision in Vehicle Technology: Land, Sea & Air. Hoboken, NJ, USA: Wiley,Feb. 2017

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

[9]. B. Yang, W. Luo, and R. Urtasun, "PIXOR: Real-time 3D object detection from point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018

[10]. J. Janai, F. Güney, A. Behl, and A. Geiger, "Computer vision for autonomous vehicles: Problems, datasets and state of the art," 2017, *arXiv:1704.05519*.

[11]. C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3D object detection from RGB-D data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018

[12]. Y. Xiang, A. Alahi, and S. Savarese, "Learning to track: Online multiobject tracking by decision making," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015

[13]. Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu, "Traffic-sign detection and classification in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016

[14]. C. Innocenti, H. Linden, G. Panahandeh, L. Svensson, and N. Mohammadiha, "Imitation learning for vision-based lane keeping assistance," in *Proc. IEEE 20th Int. Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2017

[15]. Z. Chen and X. Huang, "End-to-end learning for lane keeping of selfdriving cars," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2017

[16]. H. M. Eraqi, M. N. Moustafa, and J. Honer, "End-to-end deep learning for steering autonomous vehicles considering temporal dependencies," in Proc. Neural Inf. Process. Syst. (NIPS) ML ITS WS, 2017

[17]. 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

[18]. E. Santana and G. Hotz, "Learning a driving simulator," 2016, *arXiv:1608.01230*

[19]. Z. Yang, Y. Zhang, J. Yu, J. Cai, and J. Luo, "End-to-end multimodal multi-task vehicle control for self-driving cars with visual perceptions," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018

[20]. A. González *et al.*, "Pedestrian detection at day/night time with visible and FIR cameras: A comparison," *Sensors*, vol. 16, no. 6, p. 820, Jun. 2016

[21]. A. Gonzalez, D. Vazquez, A. M. Lopez, and J. Amores, "On-board object detection: Multicue, multimodal, and multiview random forest of local experts," *IEEE Trans. Cybern.*, vol. 47, no. 11, pp. 3980–3990, Nov. 2017.

[22]. C. Premebida, J. Carreira, J. Batista, and U. Nunes, "Pedestrian detection combining RGB and dense LIDAR data," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep. 2014

[23]. I. Sobh *et al.*, "End-to-end multi-modal sensors fusion system for urban automated driving," in *Proc. Neural Inf. Process. Syst. (NIPS) MLITS WS*, 2018

[24]. B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robot. Auto. Syst.*, vol. 57, no. 5, pp. 469–483, May 2009.

[25]. H. Xu, Y. Gao, F. Yu, and T. Darrell, "End-to-end learning of driving models from large-scale video datasets," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017

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

[27]. L. Schneider *et al.*, "Multimodal neural networks: RGB-D for semantic segmentation and object detection," in *Proc. Scandin. Conf. Image Anal. (SCIA)*, 2017

[28]. F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2016

[29]. Vineet Kosaraju, Amir Sadeghian, Roberto Mart´ın-Mart´ın, Ian D. Reid, Hamid Rezatofighi, and Silvio Savarese. Socialbigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. In Advances in Neural Information Processing Systems (NeurIPS), 2019.

[30]. Ming Liang, Bin Yang, ShenlongWang, and Raquel Urtasun. Deep continuous fusion for multi-sensor 3d object detection. In Proc. of the European Conf. on Computer Vision (ECCV), 2018.

[31]. Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In Proc. Conf. on Robot Learning (CoRL), 2017.

[32]. Boyang Deng, JP Lewis, Timothy Jeruzalski, Gerard Pons- Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Neural articulated shape approximation. In Proc. of the European Conf. on Computer Vision (ECCV), 2020.

[33]. Boyang Deng, Kyle Genova, Soroosh Yazdani, Sofien Bouaziz, Geoffrey Hinton, and Andrea Tagliasacchi. Cvxnet: Learnable convex decomposition. In Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2020.

[34]. Harm de Vries, Florian Strub, J´er´emie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C. Courville. Modulating early visual processing by language. In Advances in Neural Information Processing Systems (NIPS), 2017.

[35]. Felipe Codevilla, Eder Santana, Antonio M. L´opez, and Adrien Gaidon. Exploring the limitations of behavior cloning for autonomous driving. In Proc. of the IEEE International Conf. on Computer Vision (ICCV), 2019.

[36]. Kyunghyun Cho, Bart van Merrienboer, C¸ aglar G¨ulc¸ehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014.

[37]. Yilun Chen, Chiyu Dong, Praveen Palanisamy, Priyantha Mudalige, Katharina Muelling, and John M. Dolan. Attention-based hierarchical deep reinforcement learning for lane change behaviors in autonomous driving. In Proc. IEEE International Conf. on Intelligent Robots and Systems (IROS), 2019.

[38]. Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2017.

[39]. Shi-tao Chen, Songyi Zhang, Jinghao Shang, Badong Chen, and Nanning Zheng. Brain inspired cognitive model with attention for self-driving cars. arXiv.org, 1702.05596, 2017.

[40]. Mark Chen, A. Radford, Jeff Wu, Heewoo Jun, Prafulla Dhariwal, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In Proc. of the International Conf. on Machine learning (ICML), 2020.

[41]. Ke Chen, Ryan Oldja, Nikolai Smolyanskiy, Stan Birchfield, Alexander Popov, David Wehr, Ibrahim Eden, and Joachim Pehserl. Mvlidarnet: Real-time multi-class scene understanding for autonomous driving using multiple views. arXiv.org, 2006.05518, 2020.

[42]. Dian Chen, Brady Zhou, Vladlen Koltun, and Philipp Kr¨ahenb¨uhl. Learning by cheating. In Proc. Conf. on Robot Learning (CoRL), 2019.

[43]. Can Chen, Luca Zanotti Fragonara, and Antonios Tsourdos.Roifusion: 3d object detection from lidar and vision. arXiv.org, 2009.04554, 2020.

[44]. Sergio Casas, Wenjie Luo, and Raquel Urtasun. Intentnet:Learning to predict intention from raw sensor data. In Proc.Conf. on Robot Learning (CoRL), 2018.

[45]. Sergio Casas, Cole Gulino, Renjie Liao, and Raquel Urtasun. Spagnn: Spatially-aware graph neural networks for relational behavior forecasting from sensor data. In Proc. IEEE International Conf. on Robotics and Automation (ICRA), 2020.

[46]. Richard Bellman. Adaptive Control Processes - A GuidedTour, volume 2045. Princeton University Press, 2015.

[47]. Aseem Behl, Kashyap Chitta, Aditya Prakash, Eshed Ohn- Bar, and Andreas Geiger. Label efficient visual abstractions for autonomous driving. In Proc. IEEE International Conf. on Intelligent Robots and Systems (IROS), 2020.

[48]. Mayank Bansal, Alex Krizhevsky, and Abhijit S. Ogale. Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst. In Proc. Robotics: Science and Systems (RSS), 2019.

[49]. Waymo open dataset: An autonomous driving dataset. https: //www.waymo.com/open, 2019.