Image Segmentation of Objects in an Unstructured Environment for Autonomous Driving Systems

Tajinderpal Toor
ttoor@torontomu.ca
Electrical and Computer Engineering Graduate Student
Toronto Metropolitan University, Toronto, Canada.

***Abstract*—The field of Autonomous vehicle systems has been rapidly growing over the years. A good Autonomous vehicle system should be able to adapt and process any environment accurately. Unstructured Road Environments are commonly seen around the world in developing countries and tend to have numerous objects on the road or in a vehicle's view. Therefore Autonomous Vehicles Systems must train on unstructured data and learn how to partition these objects into different classes or image segments. Based on the Literature Review of the field it was found that this is an image segmentation problem. Furthermore, research has shown that using a U-Net architecture to identify and segment objects in an unstructured environment can be done with high accuracy. The U-Net architecture was able to identify segments within images and perform a pixel by pixel classification.**

***Index Terms*—U-Net, Image Segmentation, Object detection, Autonomous Vehicles, Unstructured Environments**

I. Introduction

For many years now, Autonomous Driving Systems have been a hot topic in many fields. An Autonomous Vehicle simply put is one that is capable of processing the information around it and traveling without human input. One of the factors that allow an Autonomous Vehicle System to perform well in a real world environment is perception. Autonomous systems should be able to extract visual information from its surroundings, be conscious and identify various objects. The Autonomous system should be able to complete this with high confidence and accuracy. Unstructured road systems or environments with no infrastructure for pedestrians and cyclists is one of the biggest challenges that autonomous driving systems face. Many third world/developing countries still have very unstructured or complex environments,in which many objects are present, not allowing trained autonomous systems to perform well. For example, by analyzing the data within the Indian Driving Data set,one can see that there are numerous objects present on the road(Vehicles,Roadside Objects,Animals,Humans,Bicycles, Rail Tracks,etc). Which means autonomous systems must classify all the objects within their surrounding before making a conscious decision. The goal of this research is to use image segmentation to classify objects that may be present within an unstructured driving environment. Specifically this paper will investigate the use of the U-Net, a fast and precise image segmentation algorithm to segment objects within the Indian Driving Data set. The Indian Driving Data set is a publicly available data set which can be found at https://idd.insaan.iiit.ac.in/.

II. Literature Review

This paper [1] proposes an idea to increase the accuracy of the U-Net, a popular semantic segmentation network used in scene understanding and object detection. The U-Net is a semantic segmented network which is U shaped and features an encoder and decoder. The encoder is used to extract relevant features from the image while the decoder outputs a segmented mask. The paper provides a method in which the encoder of the U-Net Model is replaced by a VGG-16 and ResNet-50 CNN architecture. By using a CNN architecture as the encoder of the U-Net, high resolution features can be extracted and the number of feature maps per layer is increased. The U-Net with VGG and ResNet were tested on the Cityscapes data set, and evaluated using their F1-Score and mIoU. The VGG network which consists of 7 sequential layers, 2 convolution layers followed by a ReLU activation function and 5 max pooling layers had a better performance as compared to the ResNet. Overall the proposed method of using ResNet or VGG had an improvement over the regular semantic segmentation networks.

A paper [2] was published by researchers at the Center of Excellence in Signal and Image Processing in India, that proposes a model for semantic segmentation in unstructured environments. Unstructured road environments are present in developing countries and feature unclear road boundaries and less adherence to traffic rules, along with many objects beside vehicles present on roads. The model presented in this paper introduces a modified U-Net Network, where the encoder is an EfficientNet. EfficientNet networks are constructed using Mobile Inverted Bottleneck Convolution Blocks (MBConv). The MBConv are inverted residual blocks, in which a narrow-wide structure is used as compared to the traditional residual blocks that are structured wide-narrow. Typically to increase the performance of a CNN, some sort of scaling is required. For CNN,scaling can be completed by upsurging the width, depth or input image resolution which requires lots of manual tuning. For this reason, the researchers in this paper chose to use an EfficientNet as the encoder of the U-Net as opposed to the typical CNN encoder. The EfficientNet features a method which uses a fixed ratio to scale width,depth and resolution. This paper focused on testing their methods using the Indian driving data set and concluded that using the EfficientNEetB7(7 MBConv blocks) with a U-Net decoder achieved the best result as compared to other semantic segmentation models.

Prior to the GoogLeNet model, choosing the filter size for the convolutional layer was mandatory.Through the introduction of the GoogLeNet and its inception block,improvements were made in object classification and detection. The idea in the GoogLeNet was to apply multiple filters and concatenate them instead of just choosing one. Combining the inception block from the GoogLeNet model and a typical U-Net, resulted in the U-net inception model. The U-net inception model used for image segmentation in autonomous driving was introduced by researchers [3] at Habib University in Pakistan and showed better results as compared to a regular U-Net. This paper proposed a model which features both a U-Net and GoogLeNet in order to perform image segmentation on the Mapillary Vistas Data set. The Mapillary Vistas Data set contains thousands of diverse sets of street level images with different scenes and various geographical locations. Therefore this paper introduced a robust model which allows image segmentation to be accurately performed regardless of the geographical location/weather conditions.

U-Net architectures are commonly known to be used for biomedical image segmentation, yet researchers [4] at the University of Haute Alsace in France proposed a real time road detection implementation of the U-Net architecture for autonomous driving. This paper tested and trained the U-net Model on the KITTI (Benchmark Validation Data set) and UHA (Local Highway Data set) before implementing it in a vehicle using the RT Maps Platform. The RT Maps platform was used to test the real time performance under various weather and lighting settings. The proposed model provided low complexity and good performance for real time road detection.The U-Net architecture used in this model consisted of an encoder,decoder and a block in between called the bottleneck.The U-Net architecture takes the input image and down samples to a lower resolution to decrease complexity. The encoder uses three blocks to down sample, where each of these blocks has two convolutions and uses a ReLU objective function. At the end of each block, a Dropout regularizer is present which allows unimportant neurons to be shut down after each layer, lowering the complexity. On the other hand the decoder uses three modules to up sample data and rebuild the image. In the decoder each block is composed of two convolutional layers to propagate features followed by up sampling layers to reconstruct features. The architecture ends with a sigmoid function, which outputs 0 or 1, so we can separate features that either belong or don't.

Perception and understanding the environment around a vehicle is one of the most important tasks for an autonomous vehicle to complete. This paper [5] proposes variations of the U-Net model to perform semantic segmentation on urban scenes images to try and understand the surroundings of an autonomous vehicle. The models created were compared to other semantic segmentation models(like. FCN-16, FCN-8, SegNet), in which the U-Net variants performed better. The first U-Net proposed was similar to the classic architecture. Like usual the U-Net has a contracting path and an expansive path.The contracting path also known as the down sampling was used to decrease the size of the image and extract features. Each down sampling step the number of feature channels is doubled, while the image size is decreased. On the other hand, the expansive path, also known as the up sampling path, was used to increase the image size and halve the number of feature channels. The paper also proposed four more variations of the U-Net architecture, the Small U-Net and Long U-Net. The Small U-Net variation featured two models, both with four times smaller feature channels as compared to the first model, while one uses a ReLU and the other a LeakyReLU activation function. Furthermore, the paper also proposed a Long U-Net with two layers added to the contracting and expansive path, but one was trained with a dropout rate of 0.5, while the other was trained with a dropout rate of 0.7.

As mentioned above a good autonomous driving system should provide accurate results, make decisions based on its environment and be conscious of its surroundings. This [6] paper proposes that one problem seen in autonomous driving is that the state of the art can be too expensive, complex or too hard to implement. Therefore this paper aims to show that the simple vanilla U-Net can produce good results for level 1 autonomous driving systems,be cost-effective yet still scalable to other levels. This paper used the Carla Simulator, which is a simulator that allows users to capture images with vehicles, pedestrians and traffic signals. In order to test the model, the researchers in this paper used mean pixel accuracy and intersection over union methods to evaluate their models. Using these metrics, it can be seen that the simple vanilla u-net was able to perform much better than other models such as the Fully Convolutional Neural Network.

Deep convolutional neural networks architectures give a good result and performance in tasks such as object detection and segmentation, which make them ideal for use in autonomous driving. This paper [7] makes a comparative study between various deep Convolutional Neural Networks trained from scratch on small data sets. This paper trains the models on the MiniCity data set which is a subset of the Cityscapes Data set. Since the model is being trained on a smaller data set, data augmentation must be used. Data augmentation is used to increase the amount of data by adding slightly modified copies of existing images to the data set. Researchers in this paper used the Cut Mix technique to augmentation the data. The Cut Mix technique will copy a class from one image and paste it into another image, through this process, the model can adapt and learn classes, even when a lack of data is present. Researchers implemented four CNN networks, which included DeepLabv3 and FCN both with ResNet-50 and ResNet-101. It was seen that the DeepLabv3 performed the best, especially the DeepLabv3 with ResNet-101, which had a mean intersection over union percentage of 0.64.

Due to unstructured road environments(road shoulders,walls,etc) and a complex driving system, it can be a challenge to accurately extract lane lines on the road. This paper [8] proposes a new detection method which improves the accuracy of extracting road lane lines using a U-Net and computer vision techniques such as canny edge detection and hough transform. In this paper the researchers presented a method which inputs the images into the U-Net to measure potential lane lines. Through canny edge detection and hough transforms, a vanishing point is found through the potential

lines indicated by the U-Net. Once the vanishing point is found, the horizontal lines passing through are the regions of interest. Finally, this paper proposed that using the state transition method using Gaussian distribution on the region of interest would allow accurate lane lines to be predicted. The final step of using the state transition method, will give an accurate result and will not wrongly detect lane lines due to road shoulders, walls or other elements present in an unstructured/complex road environment.

When using semantic segmentation, objects are detected using edge detection and pixel by pixel classification. When using this method, edges detected are often blurry and rough which can lead to inaccurate results. On the other hand, object detection represents an object using a rectangular box, which means objects may be misclassified. Therefore Researchers [9] at the National Cheng Kung University in Taiwan presented a method in which object detection when combined with semantic segmentation can provide accurate results in autonomous driving. The proposed network architecture has three parts: base network, segmentation decoder, and detection decoder. For the base network, researchers chose to use a fully convolutional network or a Single Shot Multi Box Detector. The base network, chosen, has seven layers, where each layer has a convolution, batch normalization and max pooling. The first layer has a filter size of 5x5 while the others have a size of 3x3 .For the segmentation decoder a MLND-Capstone deconvolution network for semantic segmentation was used, which has 7 convolution layers, 3 max pooling, 3 up sampling layers and 7 transposed convolution layers. In order to test the model, a unity engine simulation was used, which created an environment with five different objects. It was seen that the proposed approach had a n accuracy of 99.46

A lot of existing autonomous driving systems are complex, computationally heavy and require high computational power. This paper [10] proposes an experiment with different semantic segmentation models for birds eye view detection of surrounding objects. Researchers at the National Institute of Technology Karnataka, proposed a paper which used Voxelization techniques to turn three dimensional Li-DAR point clouds into two dimensional RGB images. This paper used the Lyft Level 5 data set to train and test its model. The Lyft data set provides point cloud data using the Li-DAR sensors on board. The architecture featured a modified U-Net, with 4 convolution layers and 3 deconvolution layers, where the input and target image is (336,336,3). The U-net will predict the object of every pixel in the bird's eye view image. The modified U-Net resulted in a mean precision score of 0.044.

III. Problem Statement

The field of Autonomous driving systems is gaining more attraction day by day. Autonomous driving systems can be beneficial to automotive safety, yet if they are not adaptable to uncommon environments, can be a harm to society. In order to develop systems that are robust and adaptable, we must train and develop our models on data sets that contain data from unstructured environments. Unstructured environments are present in third world or developing countries and are ones in which there are numerous other objects present in a vehicle's view. Autonomous driving systems should be able to accurately detect these objects and classify them accordingly so that a confident decision can be made. Many autonomous driving system models use data that contains structured environments to train their data, which may make the model non robust in certain environments. The problem now becomes that unstructured environments tend to have numerous different objects. Furthermore this means that each object in a frame would need to be segmented and a pixel by pixel classification would need to be completed.

IV. Dataset

The dataset used to investigate semantic segmentation and U-Net architectures was the Indian Driving Dataset. The full dataset consists of 10,000 images with 34 classes collected from 182 different road scenes. For this paper a subset of the Indian Driving Dataset, IDD Lite was used. This dataset features 8 classes with 1408 training images and 204 validation images.

V. U-Net

The U-Net is a convolutional neural network that was initially developed for use in biomedical image segmentation. Yet shows promising performance in other fields such as autonomous driving. The architecture features a contracting and expansive path which are symmetric and give the architecture a U-shape.

The U-Net architecture can be broken down into two main components, encoder, and decoder blocks. The encoder blocks are comprised of two unpadded convolutions which have a 3x3 kernel size and a stride of 1. The result of these convolutions is followed by a rectified linear unit activation function and then 2x2 max pooling operation with a stride of 2 before being fed to the following encoder block. One thing to note is that through each encoder block the number of feature channels doubles, while the image size is halved. The encoder is also known as the contracting path and uses the above-mentioned process to down sample and extract meaningful features from the image.

Now that the image has been down sampled and features extracted, the U-Net must sample and reconstruct the image so that an accurate result can be predicted. This occurs on the expanding path, also known as the decoder. Like the encoder, the decoder can also be broken down into blocks. Each decoder block is comprised of two 3x3 convolutions, concatenation, and an up-convolution. Through skip connections the result of a corresponding encoder block is concatenated with the up convolution of the previous block. This acts as the input to the first convolution in our decoder block. Through this process the feature channels is halved through each block, while the image resolution is doubled, reconstructing the image with segmentation.

VI. Proposed Architectures

U-Net 1: The first design took an image of size 126*256 and processed them through various encoder and decoder blocks. The encoder blocks had feature channels of

Layer (type)	Output Shape	Param #	Connected
input_5 (InputLayer)	[(None, 128, 256, 3)]	0	[]
conv2d_76 (Conv2D)	(None, 128, 256, 16)	448	['input_
dropout_36 (Dropout)	(None, 128, 256, 16)	0	['conv2d
conv2d_77 (Conv2D)	(None, 128, 256, 16	2320	['dropou

Fig. 1. Encoder Block Structrue U-Net

size:16,32,64,128,256. This design also featured a dropout layer, following a convolution. The model structure is shown below.

U-Net 2: The second U-Net investigated in this paper, slightly modifies the first U-Net model by expanding the feature channels. Instead of having the encoder block start at a feature channel of 16 and double through each block, we are not initializing the feature channels to 64 in the first block. This model still features dropout layers to help control over fitting.

U-Net with ResNet50 Encoder 3: The third model investigated in this paper is an ensemble of both the U-Net architecture and a ResNet50. The ResNet architecture features pre-trained weights set based on its performance on the ImageNet dataset, instead of random initialized parameters. The Segmentation Models library offered by python was used to create and test the model on the data set.

VII. Preprocessing

Before training and testing the data, the dataset needed to be preprocessed to return the most optimal results. As mentioned above, the dataset contains 8 classes that need to ne segmented. The values of these classes are [0,1,2,3,4,5,6,255]. Each of these values represent a class, therefore this would mean the last value represents the 255th class, which is not present in our dataset. Therefore, we needed to change every pixel with a value of 255 to 7, representing the 8th class to be segmented. After the 8th class was correctly added, both the image and masks were augmented. Each image and mask in the training and validation set were flipped horizontally left to right. The pictures in the dataset, were initially set to a size of 320x227, which posed a problem for the architectures proposed. As previously mentioned, at each encoder block the image size is halved, while with the decoder blocks the image size is doubled to reconstruct the image. This would mean an input image size that is divisible by two would be needed, ideally each dimension would be a power of 2. According to this the images were resized to 128x256.

VIII. Results

Unet 1 was initially trained on the dataset using a batch size of 32 and 100 epochs. The accuracy and loss for both

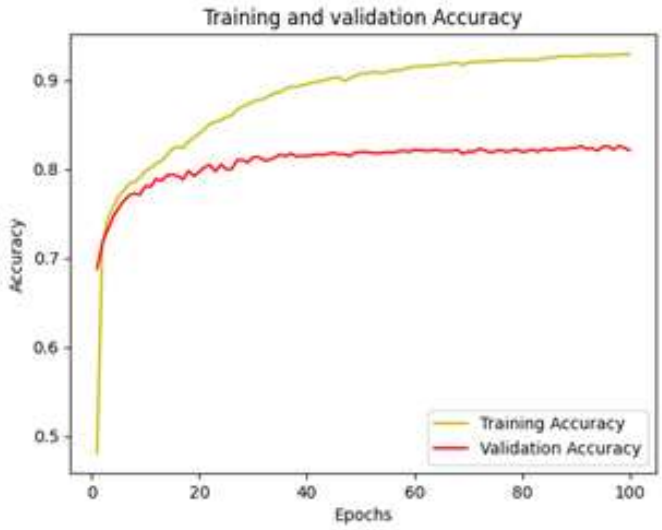

Fig. 2. Training and Validation Accuracy of U-Net 1

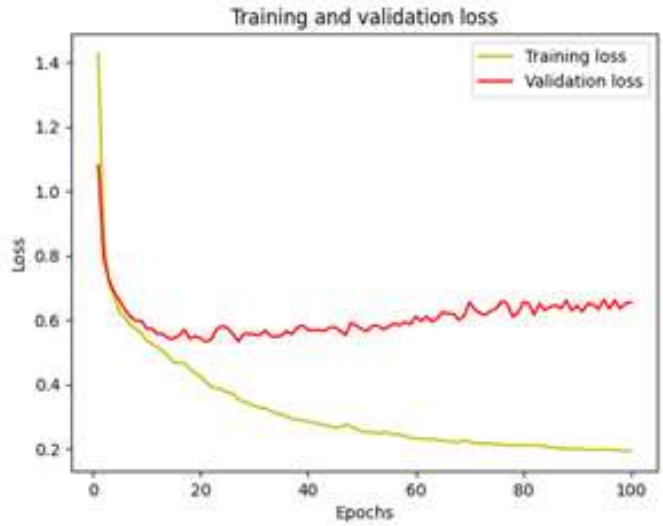

Fig. 3. Training and Validation Loss of U-Net 1

the validation and test set at each epoch can be seen in the figures below. Furthermore, figure three shows the predicted mask that was created using the U-Net 1 architecture.

The initial training performed decently and resulted in an IoU of 0.476. The model was reloaded and once again trained, from the previous values at the last epoch. The re-training was performed on a batch size of 32 over 100 epochs. The results can be seen in the figure below.

Following the completion of U-Net 1, U-Net 2 was trained.

Fig. 4. Predicted Mask from Initial Training

Fig. 5. Predicted Mask after re training model

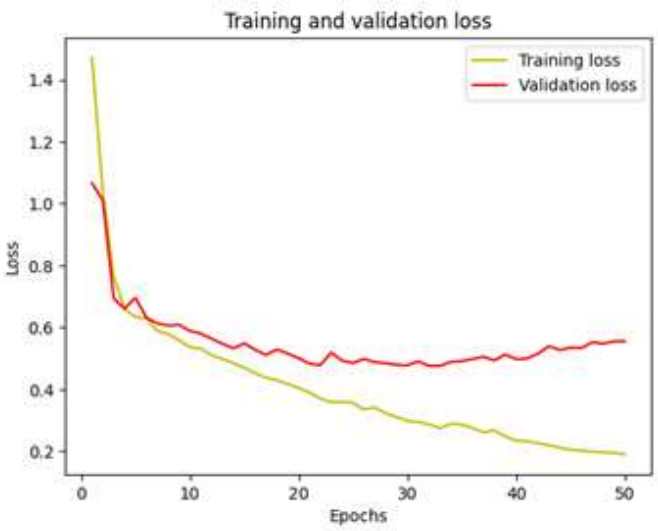

Fig. 6. Training and Validation Loss of U-Net 1

U-Net 2 was trained over 50 epochs and a batch size of 32. The loss and accuracy calculated through each epoch on the training and validation set can be seen in the figure below.

The image below, shows the predicted mask calculated using U-Net 2 for a specific instance of the test set. For this instance an IoU of 0.498 was calculated for the prediction which is an increase from our previous configuration.

Finally the U-Net with ResNet Encoder was trained. The results of this architecture can be seen in the figures below.

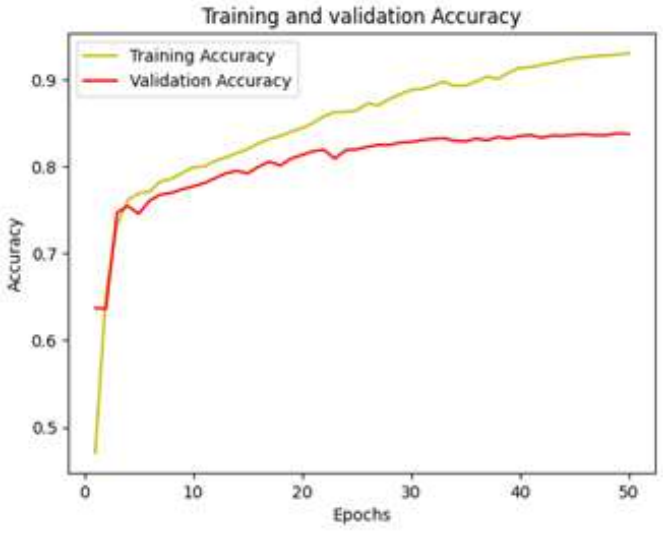

Fig. 7. Training and Validation Accuracy of U-Net 2

Fig. 8. Predicted Mask of U-Net 2

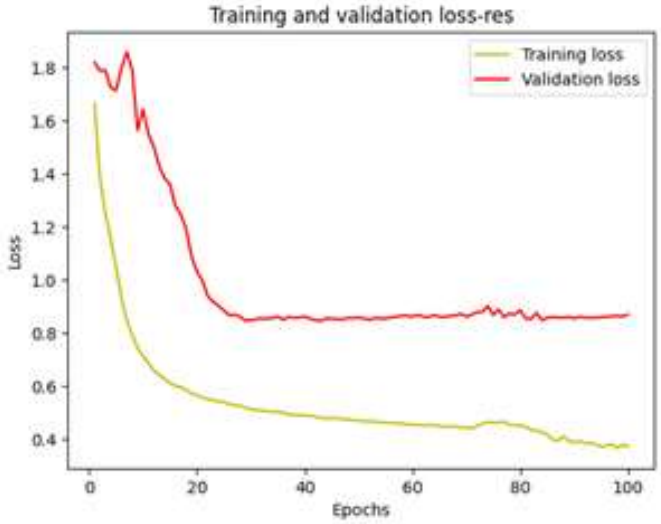

Fig. 9. Training and Validation Loss of U-Net with ResNet50 Encoder

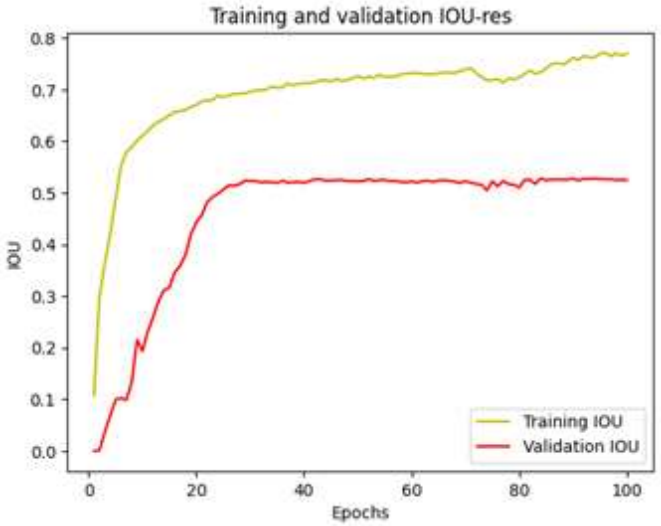

Fig. 10. Training and Validation IoU of U-Net with ResNet50 Encoder

Fig. 11. Predicted Mask of U-Net with ResNet50 Encoder

Model	Highest Accuracy	mIoU over Validation Set
UNET 1 (Design 1)	82.97	0.448
UNET 2 (Design 2)	83.72	0.466
UNET with ResNet50 Encoder	83.67	0.488

Fig. 12. Summary of Results

REFERENCES

[1] T. Sugirtha and M. Sridevi, "Semantic Segmentation using Modified U-Net for Autonomous Driving," 2022 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), Toronto, ON, Canada, 2022, pp. 1-7, doi: 10.1109/IEMTRONICS55184.2022.9795710.

[2] B. Baheti, S. Innani, S. Gajre and S. Talbar, "Eff-UNet: A Novel Architecture for Semantic Segmentation in Unstructured Environment," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 2020, pp. 1473-1481, doi: 10.1109/CVPRW50498.2020.00187.

[3] S. M. F. Hussain, S. M. Hamza and A. Samad, "Image Segmentation for Autonomous Driving Using U-Net Inception," 2022 7th International Conference on Signal and Image Processing (ICSIP), Suzhou, China, 2022, pp. 426-429, doi: 10.1109/ICSIP55141.2022.9885809.

[4] D. -V. Giurgi, T. Josso-Laurain, M. Devanne and J. -P. Lauffenburger, "Real-time road detection implementation of UNet architecture for autonomous driving," 2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP), Nafplio, Greece, 2022, pp. 1-5, doi: 10.1109/IVMSP54334.2022.9816237.

[5] H. -H. Jebamikyous and R. Kashef, "Deep Learning-Based Semantic Segmentation in Autonomous Driving," 2021 IEEE 23rd Int Conf on High Performance Computing Communications; 7th Int Conf on Data Science Systems; 19th Int Conf on Smart City; 7th Int Conf on Dependability in Sensor, Cloud Big Data Systems Application (HPCC/DSS/SmartCity/DependSys), Haikou, Hainan, China, 2021, pp. 1367-1373, doi: 10.1109/HPCC-DSS-SmartCity-DependSys53884.2021.00206.

[6] N. Darapaneni et al., "Autonomous Car Driving Using Deep Learning," 2021 2nd International Conference on Secure Cyber Computing and Communications (ICSCCC), Jalandhar, India, 2021, pp. 29-33, doi: 10.1109/ICSCCC51823.2021.9478090.

[7] A. Kherraki, M. Maqbool and R. El Ouazzani, "Traffic Scene Semantic Segmentation by Using Several Deep Convolutional Neural Networks," 2021 3rd IEEE Middle East and North Africa COMMunications Conference (MENACOMM), Agadir, Morocco, 2021, pp. 1-6, doi: 10.1109/MENACOMM50742.2021.9678270.

[8] H. Hou, P. Guo, B. Zheng and J. Wang, "An Effective Method for Lane Detection in Complex Situations," 2021 9th International Symposium on Next Generation Electronics (ISNE), Changsha, China, 2021, pp. 1-4, doi: 10.1109/ISNE48910.2021.9493597.

[9] Y. -H. Tseng and S. -S. Jan, "Combination of computer vision detection and segmentation for autonomous driving," 2018 IEEE/ION Position, Location and Navigation Symposium (PLANS), Monterey, CA, USA, 2018, pp. 1047-1052, doi: 10.1109/PLANS.2018.8373485.

[10] A. N. U, N. Naganure and S. K. S, "BEV Detection and Localisation using Semantic Segmentation in Autonomous Car Driving Systems," 2021 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), Bangalore, India, 2021, pp. 1-6, doi: 10.1109/CONECCT52877.2021.9622702