KLE Technological University

Creating Value
Leveraging Knowledge

School of
Electronics and Communication Engineering

Mini Project Report

on

ROAD SCENE SEMANTIC SEGMENTATION USING UNET AND PSP-NET

By:

1. **Sheetal Lamani** USN: 01FE21BEC401
2. **Pavan Netrakar** USN: 01FE21BEC402
3. **Kartik Malaganavar** USN: 01FE21BEC403
4. **Afaaqahamed Jamadar** USN: 01FE21BEC409

Semester: V, 2022-2023

Under the Guidance of

Prof. Bhagyashree Kinnal

K.L.E SOCIETY'S
KLE Technological University,
HUBBALLI-580031
2022-2023

SCHOOL OF ELECTRONICS AND COMMUNICATION ENGINEERING

CERTIFICATE

This is to certify that project entitled **" Road Scene Semantic Segmentation using UNet and PSP-Net"** is a bonafide work carried out by the student team of **"Sheetal Lamani 01FE21BEC401, Pavan G Netrakar 01FE21BEC402, Kartik M Malaganavar 01FE21BEC403, Afaaqahamed N Jamadar '01FE21BEC409"**. The project report has been approved as it satisfies the requirements with respect to the mini project work prescribed by the university curriculum for BE (V Semester) in School of Electronics and Communication Engineering of KLE Technological University for the academic year 2022-2023.

Bhagyashree Kinnal	**Nalini C. Iyer**	**N. H. Ayachit**
Guide	**Head of School**	**Registrar**

External Viva:

Name of Examiners	**Signature with date**
1.	
2.	

ACKNOWLEDGMENT

The sense of accomplishment that comes with having completed the Semantic Segmentation for Road scene segmentation using UNet and PSP-Net would be incomplete if we didn't mention the names of the people who helped us complete it because of their clear guidance, support, and motivation.

We are grateful to our revered institute, KLE Technological University, Hubballi, for providing us with the opportunity to realise a long-held dream of reaching the top.
We express the deep sense of appreciation and Obeisance towards our Head of School of Electronics and Communication, Dr. Nalini C. Iyer for giving the motivation and direction required for taking this industrial project to its completion. We sincerely thank our guide Prof. Heera Wali for his consistent support and suggestions.

We also thank the complete Autonomous Electric Vehicle (AEV) team for their guidance and support. Finally, we would like to thank all those who either specifically or in a indirect way made a difference in this project. We too offer profound appreciation to our guardians who have acknowledged, encouraged and helped in our endeavor.

-Sheetal Lamani, Pavan G N, Kartik M, Afaaqahamed N J

ABSTRACT

Scene understanding of urban streets is a crucial component in perception task of autonomous driving application.

Semantic segmentation has been extensively used in scene understanding which further provides assistance in subsequent autonomous driving tasks like object detection, path planning, and motion control. But accurate semantic segmentation is a challenging task in computer vision. UNet and PSP-net are the popular semantic segmentation network used for segmentation task.

In this paper, we improve the accuracy of the UNet model and the PSP-net. We are using dataset of images with traffic and other weather conditions .The analyzed data shows traffic in busy areas, and a balanced data-set allows the model to be better trained. The trained UNet and the PSP-net model.

We are hoping the proposed methodology yields better results.

Contents

Chapter 1

Introduction

In the early days of computer vision- things, or countable objects like humans, animals, and equipment, garnered the lion's share of attention. In order to understand systems that detect objects, particularly amorphous regions of similar material or texture, like grass, the sky, or a road, it is important to question the rationality of this tendency. The split of visual recognition tasks and the unique algorithms created for stuff and thing tasks both reflect the ongoing divide between stuff and things.
Inquiring about the wisdom of this pattern, The significance of researching systems that can identify objects, particularly amorphous regions made of similar materials or textures like grass, sky, or roads. The split of visual recognition tasks into stuff and thing tasks as well as the specialized algorithms created for these tasks both reflect the ongoing conflict between stuff and things.
The definition of this task is as simple as giving each pixel in an image a class label (note that semantic segmentation treats things as stuff). In contrast, the task of object detection or instance segmentation is often how studying things is defined.

Therefore Semantic Segmentation is an important part of image object detection task. It is the understanding of an image at pixel level.

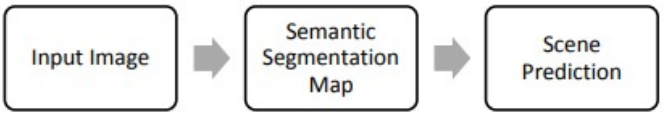

Figure 1.1: Block diagram for semantic segmentation

The above figure 1.1 shows the block diagram of the basic semantic

segmentation steps. The input image is passed to the segmentation block, where it is segmented into several classes, and later the image is mapped with a layer that separates the objects in the image and traces where they are to be placed in the image.

The majority of semantic segmentation techniques require each pixel in the image to have an associated object label. Every pixel is where the forecast is made. The forecast concerns more than just the class, as well as the limitations of each form of object. It is an approach for interpreting the look of the items and their dividing margins in the picture through pixel-by-pixel semantic segmentation. The outcome also depicts how all items in a single image—such as the sky, the terrain, and any plants—are related spatially.

1.1 Motivation

The most debated advancement in the automotive sector is self-driving technology, and many automakers are developing independent driving features for future models of their vehicles. The primary objective of this project is to address the issue of increased vehicle security, which is currently a top priority for the automotive industry.
Data show that 94% of accidents are caused by human mistake. However,many vehicles with self-driving technology avoided those measured data. Compared to human control of any vehicle, it offers greater driving security. According to the perspectives of both human and self-driving car firms, who served as the project's primary sources of inspiration, the level of importance of this project increases in the future when the majority of self-driving vehicles are electric or hybrids.

1.2 Need Statement

The need is expressed as such for the following topic: The foundation for knowing a vehicle road scenario is a precise and dependable image segmentation. Neural network models with a deep-learning foundation are used for semantic segmentation. Analyzes the proper fit for the model using a pre-scaled data set that consists of test data and multi-class data to see how well the neural network works on data that it has not previously worked with. For our project, the goal is to separate the image more

accurately and precisely compared to other neural network models and, after gathering enough data and adjusting parameters, to statistically and analytically minimise the errors from various ways.

1.3 Objectives

1. To understand how semantic segmentation might assist autonomous vehicles in accurately detecting people, vehicles, and other objects.
2. A review of the literature to determine the advantages and disadvantages of various neural network models
3. To comprehend the construction and training of the models for precise road scene segregation using a pre-scaled data set.

1.4 Literature survey

.[1]A review on deep learning techniques applied to semantic segmentation By lberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, and Jose Garcia-Rodriguez.The Model for Semantic Segmentation Based on DNN Most classifiers only compute for one category when using standard semantic segmentation techniques. When there are too many classes, it will not only result in a lot of redundancy but also have an impact on the model's performance. The subject of picture semantic segmentation has advanced significantly since the deep learning-based neural networks like DeepLabv1 and FCN were suggested in 2014. These models considerably enhance the final outcomes while being unable to anticipate all target classes directly

.[10] Pyramid scene parsing network By Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia.
The pyramid scene parsing network is a multiscale network (PSPNet). It makes use of the pyramid pooling module to advance the area of semantic segmentation. To more precisely split the scene, learn about the broader context of the scene. PSPNet has performed well in the most latest semantic segmentation ranking tables. PSPNet retrieves a baseline prior and provides semantic segmentation with additional context data. It has a significantly lower semantic segmentation error margin than FCN. The size of the receptive field indirectly affects how much picture context information is used in multilayered convolutional neural networks. By using hole convolution and feature map addition, ResNet successfully expands the receptive field; however, when the level and network level are raised, the perceptron becomes less effective.

.[4] UNet: Convolutional networks for biomedical image segmentation By Olaf Ronneberger, Philipp Fischer, and Thomas Brox. Two paths are present in the architecture. The contraction path, also known as the encoder, is the first path and it is used to record the context of the image. The encoder is simply a conventional stack of max pooling and convolutional layers. The second path, sometimes known as the decoder, is a symmetric expanding path that enables exact localisation using transposed convolutions. Because it only has Convolutional layers and no Dense layers, it is an edge to edge fully convolutional network (FCN), allowing it to process

images of any size.

.[3] Fully convolutional networks for semantic segmentation By onathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks (FCN) replace CNN's full connection layer for operation, which is the primary distinction between FCN and convolutional neural networks (CNN). The number of input layer neurons has no bearing on how the FCN functions because complete convolution is used. The local connection convolution layer may take in photos of various sizes as input. However, it is not necessary for the training picture to have the same size as the test image.From The ref .[5] The CNN network's pooling function lowers the feature map's resolution, which is particularly useful for image classification jobs because their main objective is to determine whether a certain class exists.

1.5 Problem statement

Road Scene Semantic Segmentation using PSP-Net and UNet

1.6 Application in Societal Context

The Advantage of semantics segmentation is to understand the scenerario. It is used in several work fields, like autonomous vehicle driving, robotics, medical pictures, and satellite images, as an initial step to achieve visual perception.

Autonomous driving depends on the data received by sensors of the surrounding setting so as to make an entire image of the driving scenario. Because the sign is incredibly made in such data, doing semantics segmentation correctly is crucial for scene understanding.

In addition, we tend to perform semantic segmentation with a high degree of precision on a short time scale so that the vehicle can properly understand the encompassing setting and basically create the proper call for each moment.

1.7 Organization of the report

Hence we are briefing about the contents in each chapter.

Chapter 1: It includes the report's opening, which details the first actions taken to comprehend the title and problem statement. It also includes a literature review that examines and comprehends concepts like deep learning and neural networks in relation to social environments.

Chapter 2: It includes the problem statement's design specifications. There is a functional block diagram and design theory involved.

Chapter 3: It addresses the problem statement's implementation. It includes details like the fundamental building blocks of the neural network models that are applied and the algorithms applied to implement the different functionalities.

Chapter 4: It summarises the outcomes of the project's initial Semantic Segmentation stage.

Chapter 5: It includes the project's conclusion section as well as the future strategy for this project.

Chapter 2

System design

In this chapter, we list the interfaces. The functions or techniques used to achieve the required output, as well as the methods used in obtaining these outputs, are mentioned below. Our project is divided into sections that involve the use of different neural network models. Hence, the workings of this system can be represented through a functional block.

2.1 Design alternatives

To accomplish semantic segmentation, PSP-Net and UNet are two comparable models that are employed.

2.1.1 Design 1:

Using PSP-Net for Semantic Segmentation:
To take advantage of the global context data's potential by accumulating context on a different-regional basis using our pyramid pooling module and the specified pyramid scene parsing network (PSP-Net).

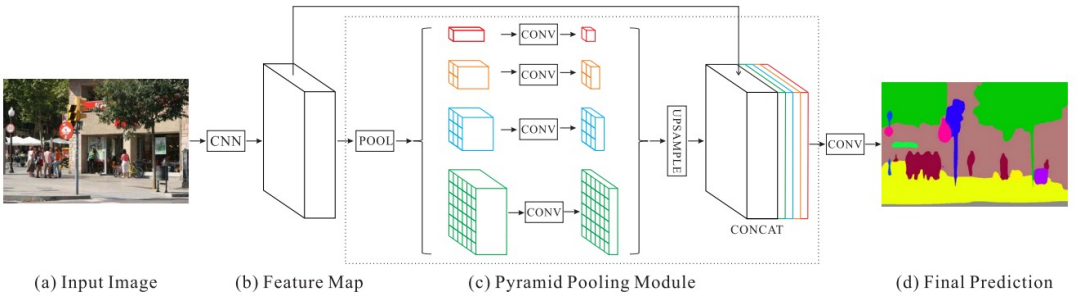

Figure 2.1

As the Figure 2.1 and the ref from [2]above shows the overview of the PSP-Net

Input Image

The network receives input images of any shape that often have size more than 256x256.

Feature Map

The feature map is extracted using a pretrained ResNet model and the dilated network technique [3, 40]. 1/8 of the input image is the size of the final feature map.

Pyramid Pooling Module

We implement the pyramid pooling module illustrated in (c). From ref [8] The pooling kernels cover all partially, and scattered areas of the image using our four-level pyramid. As the global prior, they are fused. In the final section of the formula, we then combine the earlier with the original feature map (c). The final prediction map is created in the convolution layer that comes after it.

Final Prediction

A single channel for all items, for example, or completely different channels for various things, is how the output layer is formed, and this is where the final prediction of categories is generated.

2.1.2 Design 2:

Using UNet for Semantic Segmentation:

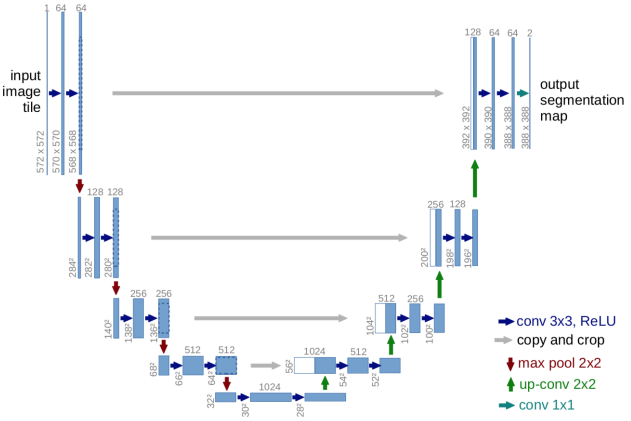

Figure 2.2

Encoder Part:

Two 3x3 convolutions ,Convolution ref from [6] must be applied repeatedly. A batch normalisation and ReLU are followed by a highly conv. In the following step, a 2x2 max pooling method is employed to reduce the abstraction dimensions. Again, we prefer to lower the abstraction levels by 0.5 while doubling the number of feature channels by a factor of two at each downsampling stage.

Decoder Part:

The feature map is resized before a 2×2 transpose convolution at each stage of the expanding route, reducing the number of feature channels in half.
In addition, we frequently combine a 3 * 3 convolutional layer with the appropriate convolution layer from the receiving path followed by ReLU. A 1×1 convolution is employed at the top layer to map the channels to the required number of categories. Also ref from [9]

Chapter 3

Implementation details

3.1 System specifications

3.1.1 Deep Learning

The deep learning field in machine learning teaches computers how to learn from experiences, just like people do. Machine learning algorithms use computer methods to "learn" information directly from data, rather than using a predefined equation as a model.

Initialization:

1. Required libraries -
   Import required python libraries and the segmentation models from pytorch framework.

2. Load Dataset -
   Eight separate classes are included in the Stanford Background Dataset that we are using. The images in the dataset's training and validation parts are either in colour or in RGB format and are composed of images. The dataset contains 715 images and is partitioned into validation and training parts in a ratio of 90:10. The dataset must also be loaded, together with the corresponding class labels and values.

3. Augmentaion and Pre-Processing -
   To perform Data visualization on dataset and one-hot encoding for each class to make it more useful for human understanding and training purpose, Also data augmentation and some pre-processing task on dataset is performed, These augmented images and masks are plotted to check if there's any unnecessary manipulation or alterations in the dataset.

4. Creating Network -
   To construct our model we are importing framework from pytorch selecting **resnet50** as our encoder, And loading weights from imagenet so this completes the loading of segmentation model with pre-trained encoder for semantic segmentation.

5. Train Network -
   Getting train and validation dataset instances and paths ready for model training and validation purpose. At the end of the all these steps, defining parameters of model training is very much important. So parameters like loss function, metrics, cuda enabling , number of epochs, optimisers, learning rate, checkpoint details with best values are tuned to achieve better accuracy at the completion of our model training.

   The weights value with best val IOU score generated from the training must be saved to perform prediction from derived weights in upcoming steps.

6. Test Network:
   Loading the best check point from saved model weights to perform prediction or segmentation on input image.To get a test image from the test dataset and passing it through the trained model will generate an output image (masked image).The quality of the output obtained depends on the accuracy of the our trained segmentation model.

7. Evaluate Performance:
   Performance evaluation plays very important role when it comes to predictability and efficiency of the model to produce the required/desired output. So we are using Dice Loss and IOU metrics for the analysis of train and validation.Then eventually Plotting the graph of Train vs Valuation.

DNN:

The user will examine the findings and select that chances the network ought to display (thosethat square measure bigger than a particular threshold, for example) before returning the projected label. every computing

is said as a layer, and complicated DNN's have various layers, thence the term "deep" networks.
DNNs are capable of modelling complex non-linear relationships. DNN architectures produce compositional models, which express the item as a layered composition of primitives. The further layers enable the combination of features from lower levels.
DNNs are feed-forward networks that transmit data from the input to the output layer without analysing it back. The DNN starts by creating a map of virtual neurons and giving its connections random whole integer values, or "weights," Increased inputs and weights provide a price between 0 and 1.
An algorithm would alter the weights if the network didn't honor a pattern correctly. As a result, the algorithm might increase the influence of specific factors until it identifies the optimal fine mathematical manipulation to fully assay the input.

CNN:

An input layer, an output layer, and numerous hidden layers make up a convolutional neural network. Convolutional layers often make up a CNN's hidden layers.
The activation function and final convolution are typically followed by extra convolutions such pooling layers, fully connected layers, and normalising layers, which are known as hidden layers because their inputs and outputs are hidden by the activation function and final convolution. In order to more accurately balance the final result, backpropagation is widely applied in the final convolution. Even while convolutions are commonly used to refer to and describe layers, this is just a convention. It is a moving dot product or cross-correlation in mathematics.

3.2 Algorithm

Semantic Segmentation using UNet and PSP-Net We have implemented the whole model on Jupyter notebook by creating our own environment to access GPU memory. The following steps show the algorithm of semantic segmentation. This method gives us the final required output. The algorithm ref is taken from [7]

Training Phase:

Step 1: Start.

Step 2: Import required python IDE and libraries .

Step 3: Divide the dataset into 90:10 ratio for training and validating respectively.

Step 4: Specify the number of classes, their names, labels and path to the train and validation folders.

Step 5: Construct the segmentation model using pytorch framework

Step 6: The metadata.csv file directs to the train and validate folder paths.

Step 7: The model is trained for appropriate batch size and epoch .

Step 8: Consider batch size 16 or 32 and epoch 60.

Step 9: Save model if a better value of IoU score is obtained. The weights of the results get stored in a separate folder.

Step 10: Best model.pth is the model obtained after training.

Testing Phase:

Step 1: Input the test image.

Step 2: The trained model will perform pixel level coloring/semantic segmentation depending on the class values as per the accuracy of the model.

Step 3: Output images are stored in a separate file.

Same steps are to be followed for the PSP-Net model

3.3 Flowchart

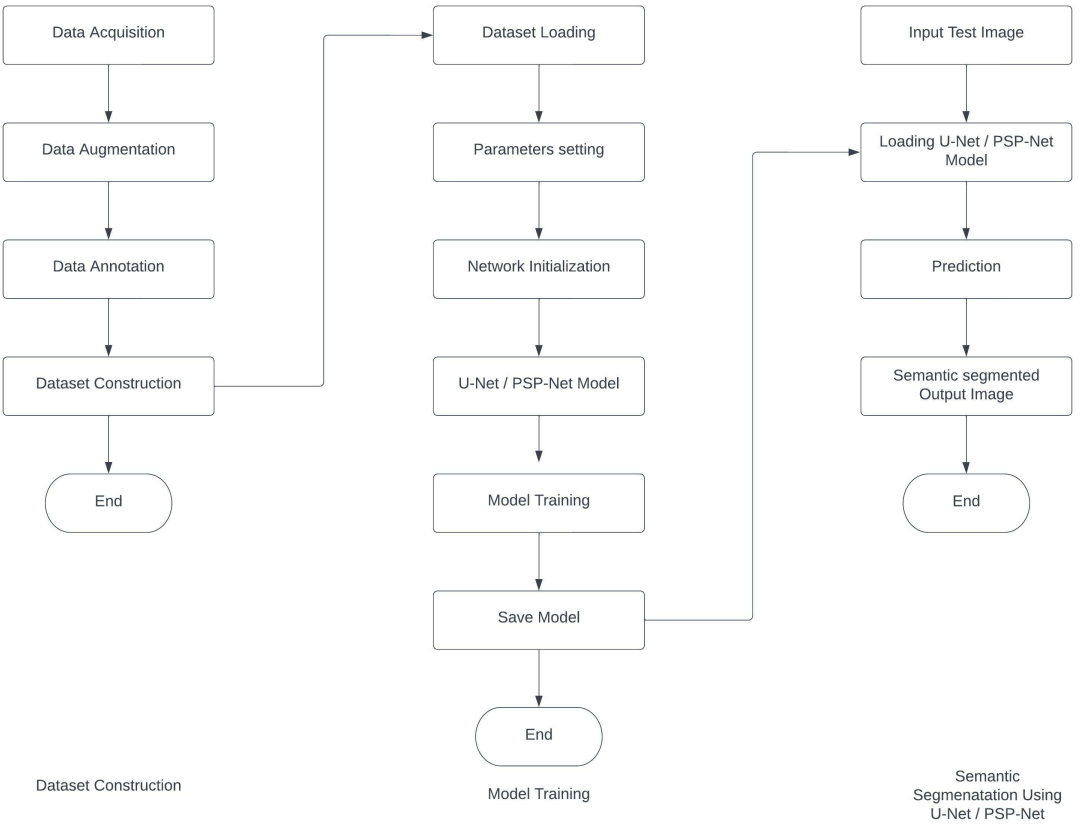

Figure 3.1

Figure **??** shows the Flow chart of the semantic segmentation using UNet/PSP-Net

Chapter 4

Results and discussions

4.1 Result Analysis

The algorithms discussed in the preceding chapter have been applied to perform semantic segmentation. The outcomes of the application are shown below.

4.1.1 Test images with it's predicted mask using UNet:

Figure 4.1

From figure 4.1, The test image is the input image that was actually used as the model's input, the ground truth is the estimated and actual segregated masked image, and the model's output is the predicted image. As a result, the similarity of the masked output can be seen when comparing it to the predicted image and the actual image. The number of classes that are separated in the input image is displayed in the various color sections.

In the figure 4.2 the second test image's accuracy is not up to the expected mark compared to the previous figure 4.1.

Figure 4.2

Figure 4.3

In the figure 4.3 The algorithm perfectly segregated the vehicles in the above image with the road.

Figure 4.4

The figure 4.4 illustrates the worst case scenario for image segmentation, in which the grass in the image and the grass reflected on the car overlap, resulting in poor segmentation.

Table showing the Training Data:

Epochs	Iou-Score	Dice Loss	
36	0.7573	0.1502	
40	0.7663	0.1438	
44	0.7838	0.1303	The model is saved at the 44th epoch because it holds the highest IoU score.
48	0.7739	0.1349	

We obtain the highest validation IoU score for the model to be saved. The table 4.1.1 above that shows the validation results of training our model.

Mean Iou-Score	Mean Dice Loss 0.7739	0.1349

In the end, the analysis of the data set yields a mean IoU score of 0.7756 and a mean dice loss of 0.1386.

IoU Score and Dice Loss plot

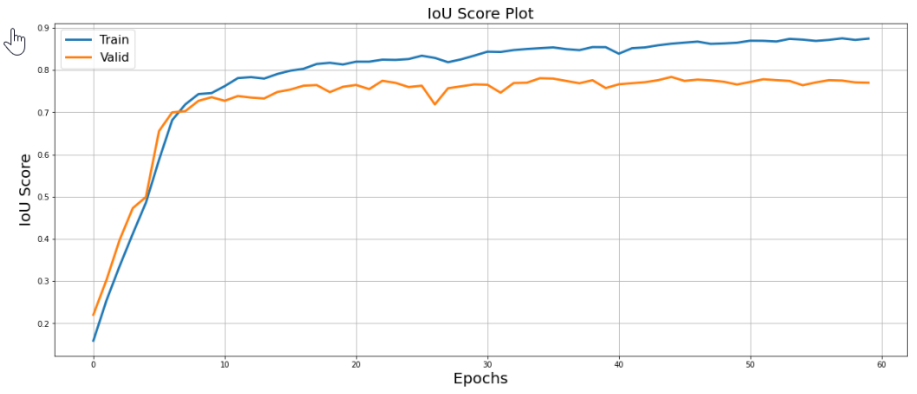

Figure 4.5

Plot 4.5 shows the correlation between epoch and IoU score, with epochs on the x-axis and IoU score on the y-axis, and the IoU score gradually rising with each epoch.

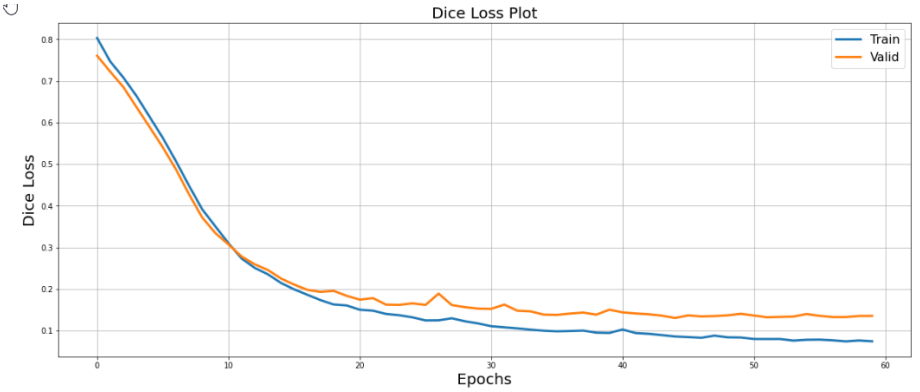

Figure 4.6

Plot 4.6 shows the correlation between epoch and Dice loss, with epochs on the x-axis and Dice sLoss on the y-axis, and the Dice loss gradually falling with each epoch.

4.1.2 Test images with it's predicted mask using PSP-Net

Figure 4.7

From figure 4.7, The test image is the input image that was actually used as the model's input, the ground truth is the estimated and actual segregated masked image, and the model's output is the predicted image. As a result, the similarity of the masked output can be seen when comparing it to the predicted image and the actual image. The number of classes that are separated in the input image is displayed in the various color sections.

In the figure 4.8 the second test image's accuracy is not up to the expected mark compared to the previous figure 4.7.

Figure 4.8

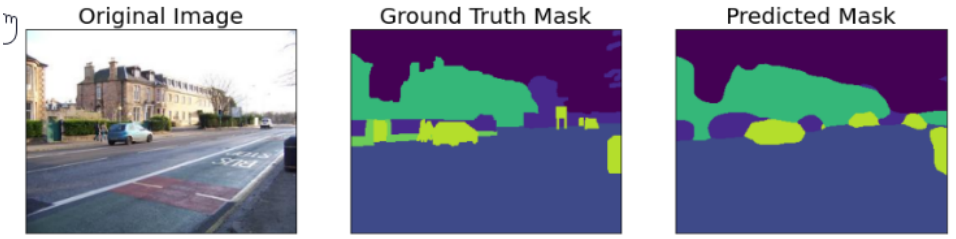

Figure 4.9

The figure 4.9 illustrates the worst case scenario for image segmentation, in which the segmentation is not precise compared to the above images

Table showing the Training Data:

Epochs	Iou-Score	Dice Loss	
92	0.7131	0.1712	
94	0.7088	0.1714	
97	0.7239	0.1630	The model is saved at the 97th epoch because it holds the highest IoU score.
98	0.7207	0.1661	

We obtain the highest validation IoU score for the model to be saved. The table 4.1.1 above that shows the validation results of training our model.

Mean Iou-Score	Mean Dice Loss 0.7290	0.1690

In the end, the analysis of the data set yields a mean IoU score of 0.7290 and a mean dice loss of 0.1690.

IoU Score and Dice Loss plot

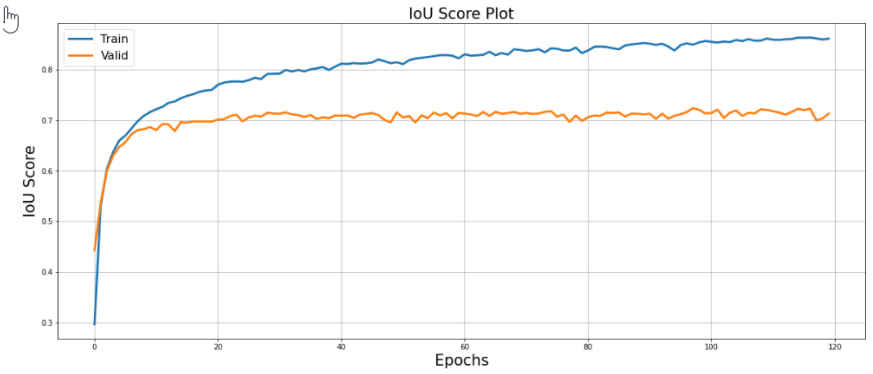

Figure 4.10

Plot 4.10 shows the correlation between epoch and IoU score, with epochs on the x-axis and IoU score on the y-axis, and the IoU score gradually rising with each epoch.

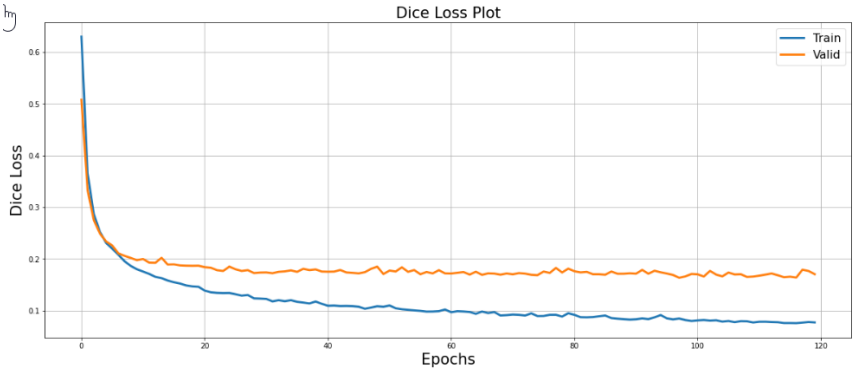

Figure 4.11

Plot 4.11 shows the correlation between epoch and Dice loss, with epochs on the x-axis and Dice Loss on the y-axis, and the Dice loss gradually falling with each epoch.

Chapter 5

Conclusions and future scope

5.1 Conclusion

This chapter summaries the conclusions we arrived after successfully implementing semantic segmentation using different models. It also tells us about the future scope of the project.

Semantic segmentation is crucial in the visual realm, and through our models, we were able to grasp the fundamentals of neural network techniques. We can gain a fundamental understanding of the evolution of semantic segmentation by putting several techniques into practise. We can also determine the structure change's trajectory. Learning is crucial for aspiring researchers. It might be useful for our upcoming studies. In the future, we intend to create a more precise and optimised implementation because of its crucial function in computer vision and machine learning.

5.2 Future scope

These models are more run on a far better computation or digital computer for better and faster computation. The long run scope of this project is to place it into action in a very bound field. As a period semantics segmentation system, It should overcome issues like weather conditions like cloudy, clear, snow and perception accuracy. Additionally, the semantic segmentation model are combined with different ADAS(Advanced Driver help Systems) modules like vehicle, sign, and road detection so as to gift an entire driver aid solution. This method, we believe, may be combined with different management systems within the future, such as lane detection improvement, a road density monitor. As a result ,more information are available, and a lot of methods are offered to options like detecting the gap between the vehicle and objects which will facilitate us to produce a scaled flight for the automobile that's safe and efficient

Bibliography

[1] Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, and Jose Garcia-Rodriguez. A review on deep learning techniques applied to semantic segmentation. *arXiv preprint arXiv:1704.06857*, 2017.

[2] Hanchao Li, Pengfei Xiong, Jie An, and Lingxue Wang. Pyramid attention network for semantic segmentation. *arXiv preprint arXiv:1805.10180*, 2018.

[3] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[5] Mennatullah Siam, Mostafa Gamal, Moemen Abdel-Razek, Senthil Yogamani, and Martin Jagersand. Rtseg: Real-time semantic segmentation comparative study. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1603–1607. IEEE, 2018.

[6] Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison Cottrell. Understanding convolution for semantic segmentation. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 1451–1460. Ieee, 2018.

[7] Yan-Ting Weng, Hsiang-Wei Chan, and Teng-Yi Huang. Automatic segmentation of brain tumor from 3d mr images using segnet, u-net, and psp-net. In *International MICCAI Brainlesion Workshop*, pages 226–233. Springer, 2020.

[8] Li-Yin Ye, Xiao-Yan Miao, Wan-Song Cai, and Wan-Jiang Xu. Medical image diagnosis of prostate tumor based on psp-net+ vgg16 deep learning network. *Computer Methods and Programs in Biomedicine*, 221:106770, 2022.

[9] Qiao Zhang, Zhipeng Cui, Xiaoguang Niu, Shijie Geng, and Yu Qiao. Image segmentation with pyramid dilated convolution based on resnet and u-net. In *International conference on neural information processing*, pages 364–372. Springer, 2017.

[10] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.