# A Comparative Analysis of Deep Semantic Segmentation Models for Self-Driving Cars

**Sarvesh Khire**
Halicioglu Data Science Institute
A59019622

**Harin Raja Radha Krishnan**
Halicioglu Data Science Institute
A59019874

## Abstract

This project aims to explore and compare the performance of three different models, namely U-Net, DeepLabV3, and DeepLab-VIT, for semantic segmentation tasks on a dataset captured from the CARLA self-driving car simulator. The dataset consists of 5000 images, each paired with labeled semantic segmentation, providing an ideal resource for training and evaluation.

By conducting this comparative analysis, the project aims to provide insights into the performance and suitability of U-Net, DeepLabV3, and DeepLab-VIT for semantic segmentation in the context of autonomous driving. Github

## 1 Introduction

Semantic segmentation plays a crucial role in computer vision applications, particularly in autonomous driving systems. The motivation behind this project lies in the increasing demand for reliable and efficient segmentation algorithms for autonomous vehicles. Accurate semantic segmentation enables vehicles to perceive and understand their surroundings, leading to improved object detection, path planning, and overall safety. By evaluating and comparing different models, we can gain insights into their strengths and weaknesses, aiding researchers and practitioners in selecting the most suitable approach for their specific application.

The project's methodology involves training the U-Net, DeepLabV3, and DeepLab-VIT models on the CARLA dataset, which comprises 5 sets of 1000 images and corresponding labeled semantic segmentations. Each model will undergo a comprehensive training process to learn the intricate relationships between input images and their corresponding semantic labels. These models leverage their unique features, such as skip connections, dilated convolutions, and vision transformers, to learn intricate spatial dependencies and capture contextual information effectively. The trained models will then be evaluated based on metrics such as pixel accuracy, intersection over union (IoU), and mean average precision (mAP).

The results of this project will provide valuable insights into the performance of U-Net, DeepLabV3, and DeepLab-VIT models in the task of semantic segmentation for autonomous driving. By comparing their accuracy, robustness, and computational efficiency, we can determine the most effective model for this specific application. The findings can contribute to the development of advanced autonomous driving systems that rely on precise and reliable scene understanding.

## 2 Related Work

Several research papers have explored the field of semantic segmentation in the context of autonomous driving, paving the way for the development of effective models for scene understanding. In this section, we highlight two papers that are closely related to our project and provide valuable insights into the application of semantic segmentation models in autonomous driving scenarios.

In the paper [1] introduced the U-Net architecture, which has become one of the most widely used models for semantic segmentation. The U-Net architecture employs an encoder-decoder structure with skip connections to capture both local and global context information. This paper demonstrated the effectiveness of U-Net in biomedical image segmentation tasks, and its principles have been successfully applied to various domains, including autonomous driving.

In the paper [4] by Chen et al. (2017) we are introduced to a prominent model in the field of semantic segmentation, known for its effective utilization of atrous convolution and fully connected conditional random fields (CRFs). This paper introduced DeepLabv2, which achieved state-of-the-art performance on several benchmark datasets. DeepLab's ability to capture multi-scale context and incorporate spatial dependencies makes it highly relevant for autonomous driving scenarios.

In the paper [7] by Zhang et al. (2021) introduces DeepLab-VIT, a novel approach that combines the power of vision transformers with the DeepLab architecture for semantic segmentation. This paper addresses the limitations of using convolutional neural networks alone by leveraging the attention mechanism of vision transformers to capture long-range dependencies and enhance feature representations. DeepLab-VIT adopts a hybrid architecture that incorporates convolutional layers in the early stages to capture local context information and vision transformers in the later stages to capture global context. By fusing the strengths of both models, DeepLab-VIT achieves state-of-the-art performance on various benchmark datasets, surpassing the performance of previous models.

In the paper [6] by Chen et al. (2018): Building upon the success of DeepLab, this paper [6] introduced DeepLabV3+, which further improved the model's performance by incorporating an encoder-decoder architecture with atrous separable convolutions. DeepLabV3+ achieved outstanding results on various challenging datasets, demonstrating its potential for accurate and efficient semantic segmentation. Its ability to handle diverse spatial resolutions and capture fine-grained details is crucial for tasks such as object detection in autonomous driving.

In the paper [8] " by Chen et al. (2016), the authors investigate the effectiveness of pre-trained ImageNet architectures for real-time semantic segmentation in the context of road-driving images. The authors explore the application of popular deep learning architectures, including VGG-16, ResNet-101, and GoogleNet, as encoders for the DeepLab framework. They demonstrate that these pre-trained architectures can provide strong feature representations for semantic segmentation tasks without the need for extensive training on specific datasets. The paper proposes an atrous spatial pyramid pooling (ASPP) module to capture multi-scale information effectively. The ASPP module employs dilated convolutions at multiple rates to gather contextual information at different scales. This approach enables accurate segmentation of objects at various sizes and maintains real-time performance.

These papers provide a strong foundation for our project, as they present influential models that have shaped the field of semantic segmentation. By referencing and understanding these works, we can leverage their insights and advancements to inform our approach and compare the performance of our proposed models, U-Net, DeepLab, and DeepLab-VIT, in the specific context of semantic segmentation for autonomous driving.

## 3 Method

Our approach for semantic segmentation is done by utilizing two popular deep learning models: U-Net and DeepLab. Semantic segmentation is a fundamental task in computer vision that involves assigning pixel-level labels to an image, enabling a fine-grained understanding of its content. To tackle this challenge, we implemented both U-Net and DeepLab architectures from scratch using the PyTorch framework. U-Net is known for its encoder-decoder structure, which allows capturing both local and global information through skip connections, while DeepLab incorporates powerful atrous convolution and dilated convolutions to capture multi-scale contextual information. By implementing these models ourselves, we gained a deeper understanding of their inner workings and customized them to suit our specific requirements. We demonstrate the effectiveness of our implemented U-Net and DeepLab models for semantic segmentation tasks, showcasing their potential for accurate and detailed image understanding.

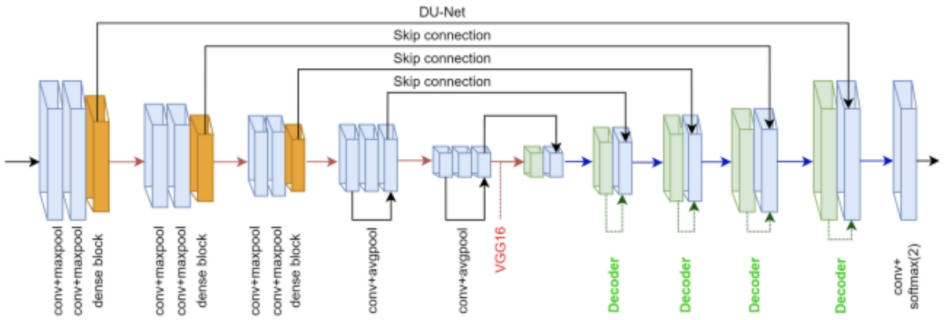



Figure 1: U-Net Architecture

### 3.1 Unet

The U-Net model [1] is a popular architecture for image segmentation tasks, particularly semantic segmentation. It was proposed by Olaf Ronneberger, Philipp Fischer, and Thomas Brox in 2015. The U-Net architecture consists of two main parts: the contracting path (encoder) and the expansive path (decoder). It gets its name from the U-shaped architecture formed by these two paths. The encoder (contracting path) and the decoder (expansive path).

#### 3.1.1 Encoder

The contracting path is responsible for capturing context and extracting features from the input image. It consists of a series of convolutional layers followed by batch norm operations. This path acts as an encoder and progressively reduces the spatial resolution while increasing the number of feature channels. The create an encoder we have created an encoding block which is basically a sequence of convolutional -> batch norm -> relu twice. Here for both the convolution operations, we have kept the kernel size as 3. There are 8 such encoding blocks in the model that serve the purpose of the encoder. In the implementation of Unet, we have also added residual connections for encoding blocks. These residual connections are added after each transpose (upsampling) operation. We will discuss them in the expansive path.

#### 3.1.2 Decoder

The expansive path (decoder) is responsible for generating a pixel-wise segmentation map. It consists of transposed convolutions along with skip connections from the encoder, which upsample the feature maps. These skip connections allow the model to utilize both low-level and high-level features, enabling the precise localization of objects in the segmentation map. The architecture of Unet is symmetrical as in it consists of a decoder that mirrors the encoder.

### 3.2 DeeplabV3

DeepLab is a family of convolutional neural network (CNN) models designed for semantic image segmentation. These models assign semantic labels to each pixel in an image, by dividing the image into different regions based on their content. There have been many versions of deep lab models, the original DeepLab model, DeepLab V2, DeepLab V3, and DeepLab V3+. In this project, we have made an attempt to write code for a smaller version of DeepLab V3. Released in 2017, DeepLab v3 [5] enhances the model's performance by employing an encoder-decoder architecture.

In the original paper, the encoder used was something similar to the Xception Net, but as it is too heavy to train with so many parameters, we wrote our own encoder using a set of convolutional layers followed by max-pooling layers. This is a stack of 10 convolutional layers, with the atrous layer as mentioned in the paper. At the end, there is a layer of transpose convolution that does the job of upsampling the feature maps and creating a segmentation map.

Deeplab V3 uses a special type of convolutional layers called as atrous convolutional layers. The atrous convolution layer, also known as dilated convolution, was introduced by Fisher Yu and Vladlen Koltun in their 2016 research paper titled "Multi-Scale Context Aggregation by Dilated

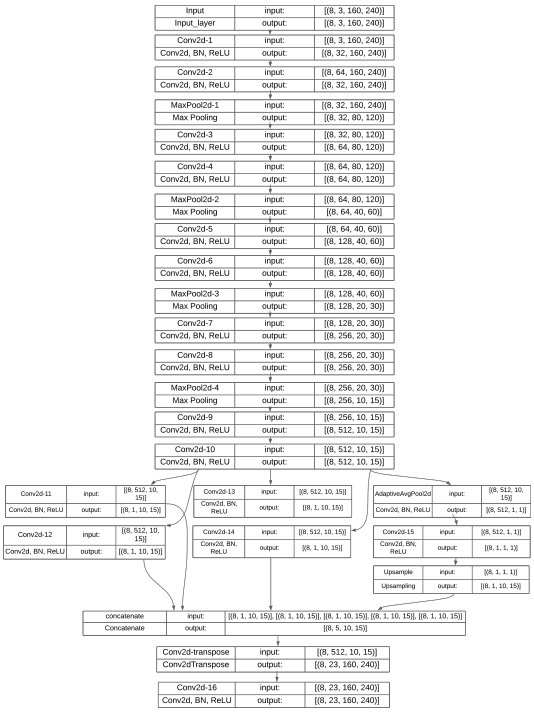



Figure 2: DeeplabV3 Architecture

Convolutions." [3] This paper proposed the use of atrous convolutions as an effective mechanism for capturing multi-scale information in convolutional neural networks (CNNs) without sacrificing spatial resolution. In Deeplab V3 we provide the output of the encoder layer to the atrous convolution layer.

### 3.3 DeepLabV3 VIT

For implementing the Deeplab VIT model, we used a similar encoder-decoder architecture with atrous layers, but here we made use of a Vision transformer as an encoder. The VisionTransformer that we used is as per [9]. We used the vit_b_16 present in the Pytorch for the encoder layer. The main idea behind using ViT is to treat an image as a sequence of patches and process them using the Transformer's self-attention mechanism.

### 3.4 Training and Testing

A dataset containing 5000 images was partitioned into a train set and a test set, with an 8:2 ratio. The train set comprised 4000 images, while the test set included 1000 images. Throughout the training process, the models utilized the Adam optimizer and employed cross-entropy loss as the objective function. The models were trained for a total of 50 epochs, with evaluations conducted every 10 epochs.

## 4 Experiments

### 4.1 Dataset

For our experiment, we utilize a dataset obtained from the Lyft Udacity Challenge, which provides images and labeled semantic segmentations captured via the CARLA self-driving car simulator. The dataset consists of five sets, each containing 1,000 images and their corresponding semantic segmentation labels.

The dataset is specifically designed for training machine learning algorithms to identify and classify semantic segmentation of various objects in an image, including cars, roads, and other relevant elements. It serves as a valuable resource for developing and evaluating models for autonomous driving applications.

The images in the dataset are typically in a standard image format, such as JPEG or PNG. Each image is associated with a corresponding label, where each pixel in the label image indicates the semantic class of the corresponding pixel in the input image. For example, different pixel values may represent classes such as cars, roads, sidewalks, buildings, etc. This pixel-level labeling allows for precise segmentation and understanding of the objects present in the scene.

In addition to the image-label pairs, the dataset may also include additional metadata or annotations that provide further information about the captured scenes, such as camera parameters, weather conditions, or vehicle poses. These additional details can be useful for improving the model's performance and analyzing the impact of various factors on semantic segmentation accuracy.

By using this dataset, we aim to train and evaluate the performance of three different models: U-Net, DeepLabV3, and DeepLab-VIT. We will analyze their ability to accurately segment and classify objects in the given images, particularly focusing on objects relevant to autonomous driving scenarios.

### 4.2 Results

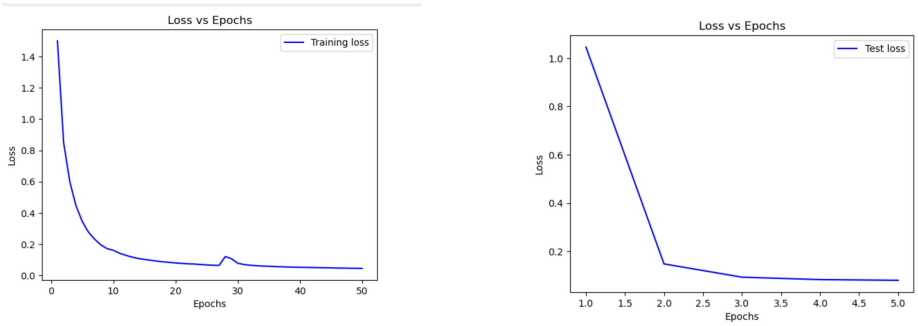
Table 1 provides a summary of the results obtained from the three models, namely U-Net, DeepLab V3, and DeepLab V3-ViT. The table presents information regarding the Train and Test Dice scores as well as the Train and Test set accuracies for each of the models.

Table 1: Model Performance

| Model | Train Dice Score | Test Dice Score | Training Accuracy | Test Accuracy |
|---|---|---|---|---|
| Unet | 0.83545 | 0.83259 | 98.34 | 97.24 |
| DeepLab V3 | 0.79730 | 0.79736 | 88.35 | 88.39 |
| DeepLab VIT | 0.77760 | 0.77762 | 83.97 | 84.07 |

From the data presented in Table 1, it is evident that the U-Net model exhibits superior performance compared to the DeepLab models. The U-Net model achieves a remarkable test accuracy of approximately 97% and a Dice score of around 0.83 for both the train and test datasets. These results highlight the effectiveness of the U-Net model in accurately predicting semantic segmentation in this context.

In Figure 3, the Train and Test loss curves of the U-Net model are depicted, illustrating the variation in loss values throughout the training process. The training loss is computed for each epoch, while the test loss is computed at intervals of every 10 epochs. Both the training and test loss exhibit a steady decline, converging to a value of approximately 0.05.

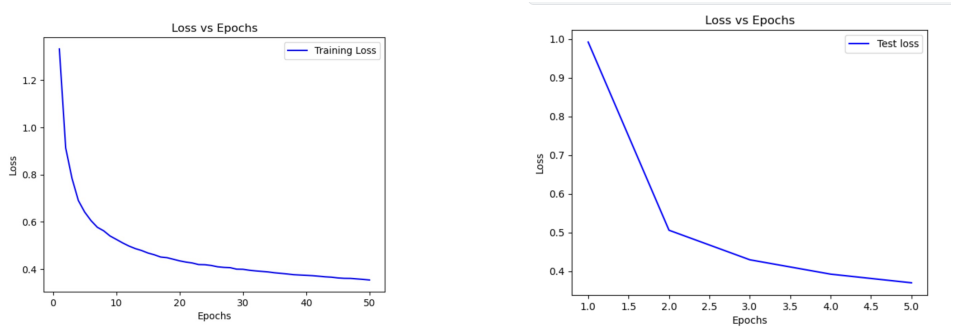



(a) Training Loss (b) Test Loss

Figure 3: U-Net Train loss and Test loss

In Figure 4, the Train and Test loss curves of the DeepLab V3 model are depicted, illustrating the variation in loss values throughout the training process. The training loss is computed for each epoch,

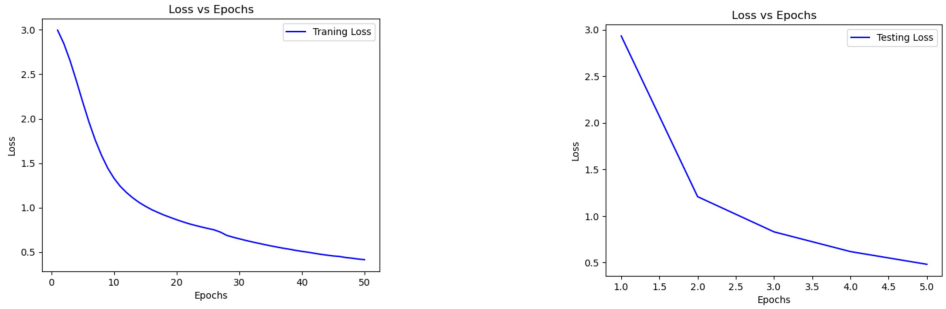
while the test loss is computed at intervals of every 10 epochs. Both the training and test loss exhibit a steady decline, converging to a value of approximately 0.38.




(a) Training Loss (b) Test Loss

Figure 4: DeepLab V3 Train loss and Test loss

In Figure 5, the Train and Test loss curves of the DeepLab V3 - ViT model are depicted, illustrating the variation in loss values throughout the training process. The training loss is computed for each epoch, while the test loss is computed at intervals of every 2 epochs. Both the training and test loss exhibit a steady decline, converging to a value of approximately 0.47.




(a) Training Loss (b) Test Loss

Figure 5: DeepLab V3 ViT Train loss and Test loss

## 5 Conclusion

In our specific case, the U-Net model has demonstrated much better performance compared to DeepLab and DeepLab-VIT in terms of segmentation accuracy and real-time processing capabilities. Several factors may contribute to these observations. Firstly, the limited size of the training dataset may have favored U-Net, as it has the ability to perform well even with a small amount of training data. Secondly, U-Net's architecture allows for dense feature extraction through the use of skip connections, enabling it to capture fine details and intricate patterns in the images. This capability may have been advantageous in our scenario. Lastly, U-Net is known for its efficient handling of object boundaries, which could have contributed to more accurate segmentations compared to the other models.

The limited training data scenario often encountered in self-driving car applications is addressed well by U-Net's architecture, which allows for information to flow directly between early and late layers, facilitating a better representation of features. Additionally, U-Net's symmetric encoder-decoder design with skip connections enables dense feature extraction at multiple scales, which is particularly useful in complex driving scenarios with objects of varying sizes and shapes. The ability of U-Net to capture both local and global contextual information further enhances its segmentation performance.

Furthermore, U-Net's effectiveness in handling object boundaries is crucial in self-driving car applications. The skip connections in the architecture help preserve fine-grained spatial information, enabling precise delineation of object boundaries.

Another advantage of U-Net is its computational efficiency, making it well-suited for real-time processing in autonomous vehicles. The relatively simpler architecture of U-Net compared to DeepLab and DeepLab-VIT contributes to its faster inference speed, allowing for timely decision-making in dynamic driving situations.

Overall, the U-Net model has shown superior performance in our study due to its ability to handle limited training data, extract dense features, handle object boundaries effectively, and support real-time processing. These characteristics make U-Net a suitable choice for semantic segmentation in self-driving car applications.

## 6 Supplementary Material

YouTube - A Comparative Analysis of Deep SemanticSegmentation Models for Self-Driving Cars
Github - GitHub Repository
Streamlit App - Streamlit Files

## References


[1] Ronneberger, O., Fischer, P., Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science(), vol 9351. Springer, Cham.

[2] Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2018 Apr;40(4):834-848. DOI: 10.1109/tpami.2017.2699184. PMID: 28463186.

[3] Fisher Yu and Vladlen Koltun. Multi-Scale Context Aggregation by Dilated Convolutions. In *International Conference on Learning Representations*, 2016.

[4] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4): 834-848, 2018.

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, Hartwig Adam. Rethinking Atrous Convolution for Semantic Image Segmentation 2017 1706.05587 arXiv

[6] Chen, LC., Zhu, Y., Papandreou, G., Schroff, F., Adam, H. (2018). Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds) Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science(), vol 11211. Springer, Cham. https://doi.org/10.1007/978-3-030-01234-2_49

[7] Sanchit Mehta, Anurag Rana, Himanshu Mehta, and Vineeth N. Balasubramanian. DeepLab-VIT: Towards Deeper and Fine-grained Visual Transformers for Semantic Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4227-4236, 2021. IEEE.

[8] Alexey Dosovitskiy and Lucas Beyer and Alexander Kolesnikov and Dirk Weissenborn and Xiaohua Zhai and Thomas Unterthiner and Mostafa Dehghani and Matthias Minderer and Georg Heigold and Sylvain Gelly and Jakob Uszkoreit and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. 2021. https://doi.org/10.48550/arXiv.2010.11929

[9] Orsic, Marin Kreso, Ivan Bevandic, Petra Segvic, Sinisa. (2019). In Defense of Pre-Trained ImageNet Architectures for Real-Time Semantic Segmentation of Road-Driving Images. 12599-12608. 10.1109/CVPR.2019.01289.