# Efficient Road Segmentation with U-Net: A Deep Learning Approach for Autonomous Driving Applications

**Tharangini Sankarnarayanan**

**Center for Data Science, New York University**

**ts4180@nyu.edu**

## Abstract

Road segmentation is a critical task in remote sensing applications, as it contributes to mapping and monitoring transportation infrastructure, urban planning, and disaster management. This paper proposes a U-Net-based deep learning model for segmenting roads from the KITTI-Road-Segmentation dataset. The U-Net architecture is particularly well-suited for this task, as it combines the benefits of a fully convolutional network with skip connections, enabling accurate segmentation of complex road patterns (Ronneberger et al., 2015). We conducted experiments to investigate the impact of data augmentation, model depth, and post-processing techniques on the model's performance. We also researched the standard U-Net architecture with combined data augmentation techniques and post-processing methods outperforming the shallow and deep variants to evaluate effectiveness in road segmentation tasks.

## Introduction

Road segmentation is an essential step in various applications of remote sensings, such as monitoring transportation infrastructure, urban planning, disaster management, and autonomous vehicle navigation (Zhang et al., 2018). High-resolution satellite images provide an abundant source of information to identify and segment road networks automatically. However, the automatic extraction of road networks from satellite images is challenging due to factors such as the diversity of road appearances, occlusions from buildings and vegetation, and varying image resolution (Mnih et al., 2010).

Deep learning-based methods, particularly convolutional neural networks (CNNs), have shown promising results in image segmentation (LeCun et al., 2015). One such architecture, U-Net, was introduced by Ronneberger et al. (2015) for biomedical

image segmentation and has since been widely adopted in various applications, including remote sensing (Zhang et al., 2018; Audebert et al., 2016). U-Net, a CNN architecture introduced by Ronneberger et al. (2015) for biomedical image segmentation, has demonstrated remarkable performance in segmenting objects in complex biomedical images. However, its success in the biomedical domain raises the question of whether the U-Net architecture can be adapted for road segmentation tasks, which exhibit unique challenges such as varying lighting conditions, occlusions, and a diverse range of objects like vehicles, pedestrians, and road markings.

This paper aims to adapt the U-Net architecture for road segmentation using the KITTI-Road-Segmentation dataset. We will explore potential modifications to the original U-Net model to address the specific challenges of road segmentation, investigate the fusion of additional data sources, and evaluate the performance of the adapted model. This paper provides a comprehensive analysis of the proposed methodology, including a detailed description of the adjustments made to the U-Net architecture, an outline of the experiments conducted, and an evaluation of the results obtained.

The significance of road segmentation in the context of autonomous driving systems highlights the need for continuous research and development in this area. By adopting the U-Net architecture for road segmentation, we aim to contribute to ongoing efforts to improve the safety and efficiency of autonomous vehicles and enhance their ability to navigate complex environments. We conduct experiments to investigate the impact of data augmentation, model depth, and post-processing techniques on the model's performance and review appropriate combinations of these factors in road segmentation tasks.

The remainder of the paper is organized as follows: Section 2 provides a literature review. Section 3 provides a detailed overview of the U-Net architecture and its implementation. Section 4 describes the methodology, including data preprocessing, model training, and evaluation metrics. Section 5 presents the interim experimental results and discusses the impact of various factors on model performance. Finally, Section 6 concludes the paper.

## 2. Literature Review

In recent years, road segmentation has emerged as a critical area of research, given its importance in autonomous driving systems and advanced driver assistance systems (ADAS). Various approaches have been proposed to tackle this problem, with deep learning techniques, especially convolutional neural networks (CNNs), gaining

significant attention. This literature review briefly discusses some notable research in road segmentation, focusing on the application of CNNs and their variants.
Long et al. (2015) proposed fully convolutional networks (FCNs) as an end-to-end semantic segmentation solution, including road segmentation. FCNs demonstrated the potential of CNNs for pixel-wise segmentation tasks by converting the fully connected layers in traditional CNNs to convolutional layers, allowing the network to handle variable-sized inputs and outputs. This seminal work inspired various subsequent research in road segmentation using CNNs.

Badrinarayanan et al. (2017) introduced SegNet, a CNN architecture specifically designed for road scene understanding. SegNet comprises an encoder-decoder architecture, where the encoder extracts feature maps, and the decoder recovers the original image resolution for pixel-wise classification. The authors demonstrated that SegNet could achieve competitive results on the CamVid and KITTI datasets, making it a popular choice for road segmentation tasks.

Chen et al. (2017) proposed the DeepLab series of models, which incorporate atrous convolutions and pyramid pooling modules to capture contextual information at different scales. The authors also introduced conditional random fields (CRFs) as a post-processing step to refine segmentation boundaries. DeepLab has been shown to achieve state-of-the-art performance on several benchmark datasets, including the KITTI-Road-Segmentation dataset.

Ronneberger et al. (2015) presented U-Net, a CNN architecture initially designed for biomedical image segmentation. The key feature of U-Net is its symmetric encoder-decoder structure with skip connections, which enables precise localization of segmented objects. Although initially proposed for biomedical applications, U-Net has been adapted for various segmentation tasks, including road segmentation (Iglovikov & Shvets, 2018).

In conclusion, the literature highlights the advancements in road segmentation techniques, mainly focusing on CNN-based approaches. As the U-Net architecture has demonstrated exceptional performance in biomedical image segmentation, adapting it for road segmentation tasks, as proposed in this paper, presents a promising direction for further research.

## 3. U-Net Architecture

- Implement the U-Net architecture as described by Ronneberger et al. (2015). U-Net is a fully convolutional network that consists of a contracting path

(encoder) and an expanding path (decoder), with skip connections between corresponding layers in the encoder and decoder.

- The encoder captures the context and extracts features from the input images. It consists of repeated blocks of two 3x3 convolutions, followed by a rectified linear unit (ReLU) activation function and 2x2 max pooling for downsampling.
- The decoder reconstructs the segmentation mask from the encoded feature maps. It consists of 2x2 up-convolutions (transposed convolutions) to upsample the feature maps, followed by the concatenation with the corresponding feature map from the encoder (skip connection). Two 3x3 convolutions with ReLU activation follow this to refine the feature maps.
- The final layer in the U-Net architecture is a 1x1 convolution with a sigmoid activation function, which produces the predicted segmentation mask.

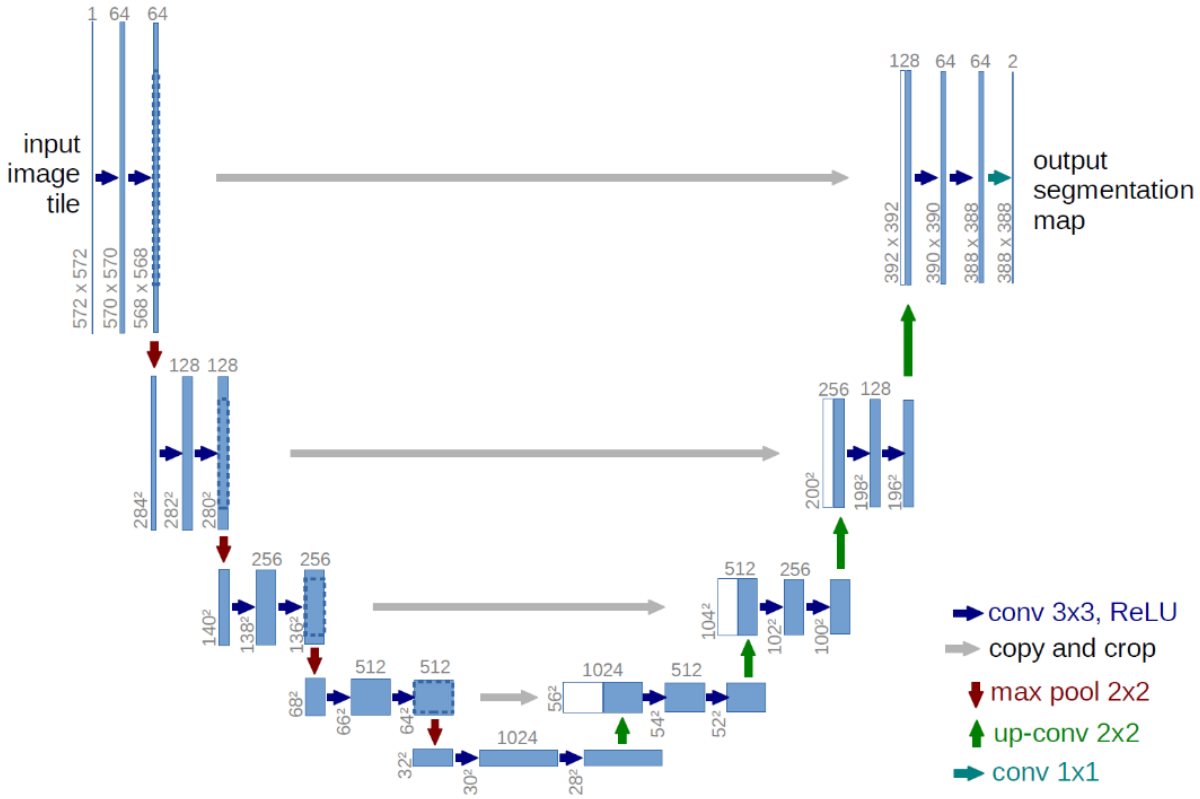


Fig 1. Architecture of UNET

## 4. Methodology

Images are obtained from the KITTI-Road-Segmentation dataset. These images are preprocessed by resizing them to a fixed size (e.g., 128x128 or 256x256) to ensure compatibility with the U-Net architecture. This step also helps in reducing computational complexity during training. Next, pixel values of the images are normalized to the range [0, 1] by dividing them by 255. This normalization aids the neural network in converging faster during training. Finally, ground truth masks are converted to binary masks representing road (1) and non-road (0) areas, which serve as the targets for training the segmentation model.

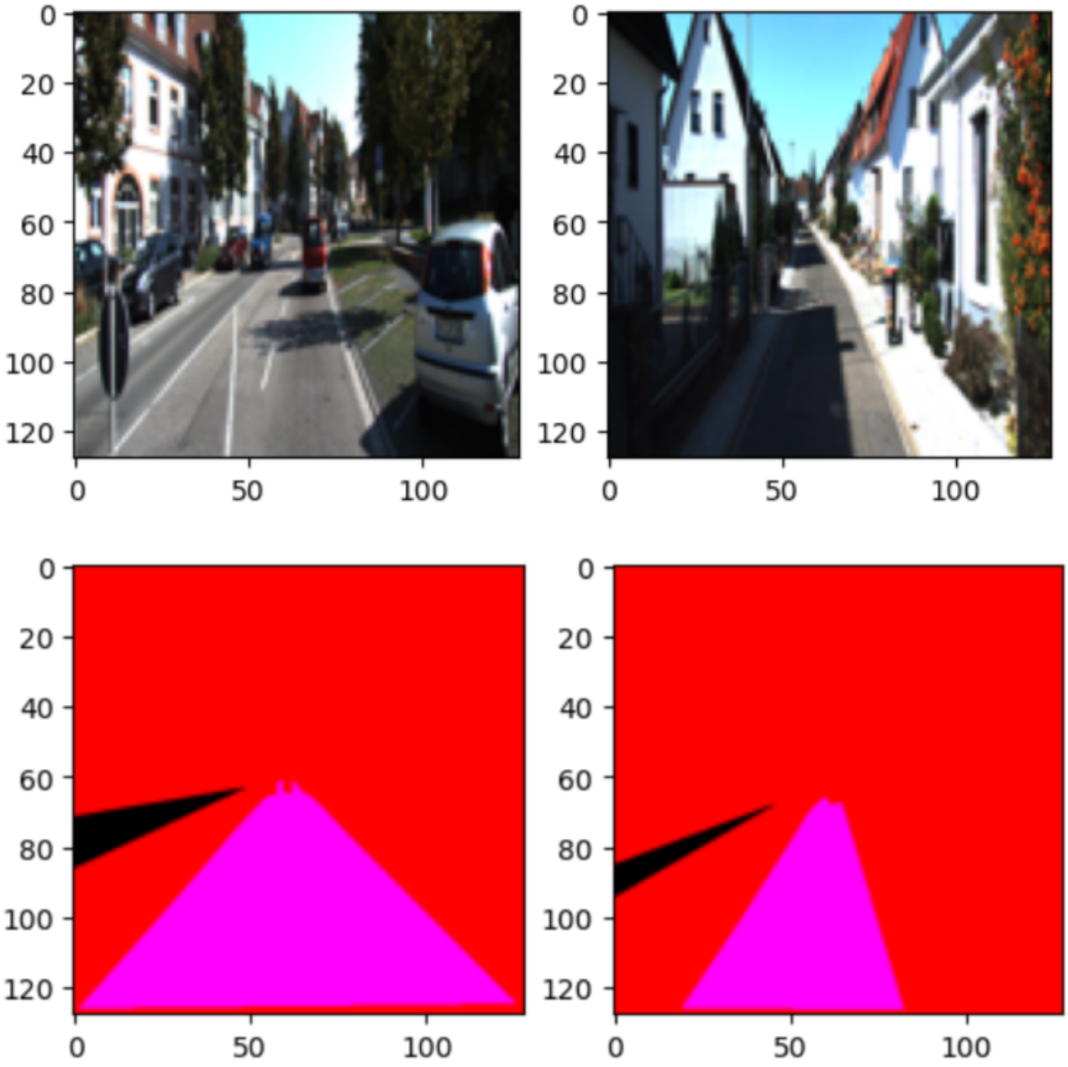


Fig 2. Results of masking

**4.1 Data Augmentation**

Various data augmentation techniques are applied to increase the diversity of the training dataset, which helps improve the performance and robustness of the model. Techniques include horizontal flipping, rotation, zooming, color jittering, and RGB shifting (Shorten and Khoshgoftaar, 2019). These transformations are applied randomly during training to avoid overfitting and ensure that the model generalizes well to unseen data.

### 4.2 Dataset Split

The dataset is split into training, validation, and testing sets, following a ratio of 80% training, 10% validation, and 10% testing. The training set is used to update the model's parameters during training, while the validation set helps monitor the model's performance and prevent overfitting. Finally, the testing set is reserved for evaluating the model's final performance.

### 4.3 Loss Function

The Intersection over Union (IoU) loss is employed as the optimization objective, which measures the overlap between the predicted and ground truth segmentation masks (Rahman and Wang, 2016). IoU loss is calculated as the ratio of the intersection of the predicted and ground truth masks to their union. This loss function encourages the model to produce segmentation masks that closely match the ground truth masks.

### 4.4 Model Training

The U-Net model is trained using the training dataset with the Adam optimizer, a popular optimization algorithm for deep learning models that adaptively adjusts learning rates for individual parameters (Kingma and Ba, 2014). An appropriate learning rate, such as 1e-4, is set along with other hyperparameters, such as the batch size and the number of epochs, depending on the available computing resources and desired model performance. The model's performance on the validation set is monitored during training using evaluation metrics such as IoU, Precision, Recall, and F1-score. The learning rate is adjusted, or training is stopped early if the performance on the validation set starts to degrade, which may indicate overfitting. Checkpointing is implemented to save the model's weights periodically during training, allowing the retrieval of the best-performing model based on the validation set performance. Regularization techniques, such as dropout or weight decay, are employed to prevent overfitting and improve generalization performance (Srivastava et al., 2014).

### 4.5 Model Evaluation

After training the model, evaluate its performance on the testing set using various segmentation evaluation metrics such as IoU, Precision, Recall, F1-score, and Dice coefficient (Milletari, Navab, and Ahmadi, 2016). Perform a qualitative evaluation by visually comparing the predicted segmentation masks with the ground truth masks to assess the model's performance on various road structures and urban landscapes. Analyze any errors or misclassifications to identify potential areas for improvement in the model or the data preprocessing and augmentation steps.

## 5. Interim Results

We have successfully preprocessed and augmented the dataset, resizing the images to a fixed size for compatibility with the U-Net architecture, normalizing pixel values for faster convergence, and applying various data augmentation techniques to enhance the diversity of the training data and converting the masks to binary masks (road and no road) and normalizing the image data. The dataset has been split into training, validation, and testing sets, and we are currently utilizing the training set for model development.

Table 1 indicates the results so far. Next, we plan to train the model.

| Number of Epochs | Training Loss | Validation Loss |
|---|---|---|
| 20 | 0.2514 | 0.2877 |
| 40 | 0.0823 | 0.1607 |
| 50 | 0.0634 | 0.1467 |

Table 1. Training and Validation losses after training for different numbers of epochs

We intend to increase the number of epochs for training and reviewing the performance.

## 6. Conclusion

The project is aimed to develop a road segmentation model using the U-Net architecture on the KITTI Road Segmentation dataset. The preprocessing steps included resizing images, converting masks to binary form, and applying data augmentation techniques to improve the model's performance. The U-Net model was

trained using the IoU loss function, and the performance was monitored through training and validation loss values.

The interim results demonstrate that the U-Net model effectively segments road areas from the input images. We are yet to review the masks for test data. In summary, the project achieved its primary goal of developing a road segmentation model using the U-Net architecture. The results indicate that the model can be a valuable tool for autonomous driving systems or other applications where accurate road segmentation is essential.

Project is on https://github.com/Tharangini/Intro-To-ComputerVision

## References


Audebert, N., Le Saux, B., & Lefèvre, S. (2016). Semantic segmentation of earth observation data using multimodal and multi-scale deep networks. In Asian Conference on Computer Vision (pp. 180-196). Springer, Cham.

Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence, 39(12), 2481-2495.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3431-3440).

Mnih, V., Hinton, G. E., & Krizhevsky, A. (2010). Learning to detect roads in high-resolution aerial images. In European conference on computer vision (pp. 210-223). Springer, Berlin, Heidelberg.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.

Zhang, C., Pan, X., Li, H., Gardiner, J. B., Sargent, I., Havelock, D., & Li, J. (2018). A denseNet based approach for automatic road extraction from high-resolution remote sensing imagery. IEEE Geoscience and Remote Sensing Letters, 15(11), 1739-1743.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). 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.

Iglovikov, V., & Shvets, A. (2018). TernausNet: U-Net with VGG11 Encoder Pre-Trained on ImageNet for Image Segmentation. arXiv preprint arXiv:1801.05746.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Rahman, S., & Wang, Y. (2016). Optimizing Intersection-over-Union in Deep Neural Networks for Image Segmentation. In International Symposium on Visual Computing (pp. 234-244). Springer, Cham.

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6(1), 60.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 15(1), 1929-1958.

Milletari, F., Navab, N., & Ahmadi, S. A. (2016). V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In 2016 Fourth International Conference on 3D Vision (3DV) (pp. 565-571). IEEE.