UNet and UNet+ResNet: A Comparative Analysis for Lane Detection in Autonomous Vehicles

Minahil Bakhtawar
University of Toronto
1008548957

Sneha Balaji
University of Toronto
1007999212

Rachel Chan
University of Toronto
1007853734

***Abstract*—Lane detection is a vital component of autonomous driving, particularly in identifying key characteristics of the surroundings such as the vehicle's relative position, direction of travel and road rules [4]. However factors such as lighting, weather conditions and road surface texture pose hurdles to this task. In recent years, deep learning approaches have shown promise in addressing these challenges, with various combinations of traditional models like ResNet and UNet being explored. [6] These methods have demonstrated success in semantic segmentation tasks, including disease detection in biomedical imaging. [7] [10]. Thus, this project compared the performance of a traditional UNet model against a UNet model with ResNet-50 encoding blocks using the TuSimple [13] dataset. It was discovered that the Res-UNet model had higher scores in recall, F1 and IoU while UNet had higher accuracy and precision scores.**

***Index Terms*—Lane detection, UNet, ResNet, autonomous driving, semantic segmentation**

I. Introduction

The surge in autonomous driving, for both its novelty and potential to reduce traffic accidents, has given rise to numerous machine learning tasks such as path planning, obstacle avoidance, and, most notably, lane line detection [1]. Lane detection identifies road lane markings, providing vehicles with vital information such as the direction of travel, their relative position, passing rules [4]. Challenges arise in lane detection due to the high variability between scenes caused by lighting, blur, weather conditions, road surface texture and more. Deep neural networks have shown great promise in lane detection due to their ability to quickly and accurately capture complex features, with leading models obtaining over 90% accuracy [3].

Lane detection is done using semantic segmentation which predicts a label for every pixel in a given image, deciding it whether or not the pixel belongs to a lane, with the number of line classes varying between datasets [2]. UNet-based segmentation networks are commonly used for this task due to their encoding-decoding architecture and skip connections that are able to capture both high and low-resolution features while retaining their spatial information, achieving upwards of 90% accuracy [3]. UNet's limitations arise from the vanishing gradient problem, which is a challenge found when training deep networks during the back propagation step of gradient descent. Back propagation is used to better fit the training data by updating a model's weights with respect to the gradient of the loss function, however as more layers are added the gradient can quickly diminish to zero, effectively stagnating the earlier layer's weights to a sub-optimal value. UNet slightly offsets this problem with the use of skip connections between its encoding and decoding arms, but suffers when trying to capture more complex scenes like the ones seen in lane detection images [5]. This requires a better understanding of the global contextual situation, while retaining long range dependencies. The primary objective of our research is to enhance the performance of semantic segmentation models for autonomous driving applications by leveraging a ResNet backbone.

ResNet was designed to address this vanishing gradient issue and has seen great success due to its residual block architecture [8]. These residual blocks place skip connections between the input of each block and its output, with the difference (residual) between the two allowing for deeper network architectures [9]. While traditional ResNet is not best suited for image segmentation tasks, with it performing best in image classification, it can be easily combined with other networks to address their vanishing gradient problem. Additionally, it improves the understanding of more abstract, hierarchical features in image segmentation which can improve UNet's understanding of the global context in lane detection [6]. Different combinations of ResNet and UNet, commonly referred to as ResUNet, have been proposed for various semantic segmentation tasks including change detection, remote-sensing, and disease detection, with particularly promising architectures utilizing a UNet segmentation framework and ResNet encoding blocks [6] [7] [10].

This project proposes the use of one such ResUNet model to perform lane detection on the TuSimple Lane Detection Dataset, which is a widely used benchmark dataset in autonomous driving research [13]. The UNet segmentation framework will be trained with ResNet-50 encoding blocks and a traditional UNet network on this dataset for comparison.

The next part of this proposal discusses related works in the Literature Survey section and goes on to describe the chosen dataset, problem statement, and system model in further detail.

II. Literature Survey

A. Robust U-Net-based Road Lane Markings Detection for Autonomous Driving [1]

Tran et. al propose a novel method for detecting lane road markings by applying a three step process using a U-Net architecture model, Hough Transform and K-means clustering. The researchers use a U-Net model to extract road lane features, then Hough Transform for edge detection and finally K-means clustering to determine which lines most closely match the original road lane marking. For the input, this study uses 512 x 512 pixel road images from a front-view camera sourced from the CARLA simulator [12] and the performance of the model was over 100 FPS with the driving simulator.

B. Identification of Road and Surrounding Obstacles using U-Net Architecture for Better Perception [2]

This research study explores semantic segmentation on road lanes and surrounding objects using U-Net architectures and images with labels from the CARLA dataset [12] with a pre-processing step of resizing the images from 600 x 800 x 3 pixels to 96 x 128 x 3 pixels. The results of 3 separate U-Net networks are concatenated and several convolution layers are applied to generate activation maps for each network's mask layer. The study achieved a training accuracy of 92.27% and a validation accuracy of 88.65%.

C. Lane line detection algorithm based on improved UNet network [3]

The researchers in this study propose using CBAM to detect road lanes in the TuSimple dataset [13] due to its ability to adaptively learn the importance weights of spatial locations and feature channels during the training process. Thus, CBAM is able to enhance the feature extraction process by learning the relationships between feature channels allowing the model to pay greater attention to feature channels relevant to lane detection. The study used the Tusimple dataset to evaluate and model and found it to perform better than the traditional U-Net model with an accuracy of 96.82% and an F1 measure of 97.14%.

D. Deep Learning Based Segmentation Approach for Automatic Lane Detection in Autonomous Vehicle [4]

In this paper, two commonly used deep learning models SegNet and U-Net are compared in their performance in semantic segmentation for lane detection tasks using mean squared error (MSE), average miss and overall computation time to assess the models' performance. The lane detection dataset with annotated images from the tuSimple website is used. Findings present that U-Net has a lower MSE of 13.58% compared to the SegNet of 16.85% and a lower average miss of 0.341 compared to SegNet's 0.124 but U-Net is found to have a longer computation time of 22.79 ms compared to SegNet's 20.45 ms.

E. Self-Attention blocks in UNet and FCN for accurate semantic segmentation of difficult object classes in autonomous driving [5]

Mousavi et. al propose the use of Self Attention blocks in UNet to improve accuracy of pedestrian detection using the cityscapes dataset since CNNs struggle to capture long-range dependencies or correlations between distant pixels in an image. Researchers use the Adam optimizer to train weights, and mIoU as the main metric to compare competing models. It is found that combining Unet with SA blocks yields improved model performance and similar results can be expected with ResNet-Unet in lane detection as ResNet's skip connections enable the model to capture and preserve fine details and spatial relationships.

F. Enhancing Change Detection in Spectral Images: Integration of UNet and ResNet Classifiers [6]

E. Brahim et al. propose combining UNet and ResNet networks to leverage both of their strengths for a remote sensing change detection (CD) task since UNet excels in image segmentation tasks due to its encoding-decoding architecture and ResNet is desirable as it fixes the vanishing gradient problem. The input to both networks was two multi-spectral NDVI images from the same location at different times, with the ensemble model averaging the output from both of these networks to achieve the final change detection. They report a final accuracy of 99.50% and an F1-score of 99.41%, which outperformed 5 state-of-the-art models with fewer false changes and a 15 to 50-second faster response time.

G. Semantic Segmentation of Venous on Deep Vein Thrombosis (DVT) Case using UNet-ResNet [7]

A. K. Hernanda et al. propose using a UNet segmentation framework with Resnet-32 contraction modules, shown in figure 3, to binarily identify Deep Vein Thrombosis (DVT) vein areas in ultrasound images. Their results show that UNet suffered from a vanishing gradient problem due to the sigmoid activation, while the ResUNet did not have the same issue. ResUNet outperformed Unet with >3% higher IoU score and lower dice loss, which were [84.50%, 0.0857] for ResUNet and [81.22%, 0.1341] for UNet. These results indicate that a ResUNet architecture can outperform UNet on an image segmentation task using gray-scale images and binary classification labels.

H. Classification of chest pneumonia from x-ray images using new architecture based on ResNet [8]

This paper discusses the utility of Resnet in the biomedical field when diagnosing chest pneumonia as the skip connections in Resnet allow for information to be retained, and an overall higher accuracy. They use the kaggle chest X-ray dataset and use Resnet-50 with Dropout regularization and ReLU to train their model. When extrapolating these findings to autonomous driving applications, the utilization of ResNet is anticipated to enhance the accuracy of object detection and scene understanding due to its capability to grasp the global context of the image, consequently benefiting lane detection tasks.

I. Semantic segmentation of unmanned aerial vehicle image based on Resnet-Unet [9]

This paper proposes a Resnet-Unet approach to solve the issue of low semantic segmentation accuracy in UAV images, which is caused by target overlap, and unbalanced distribution of target objects. A Resnet-50 encoder backbone is used with UNet on the Aeroscapes dataset, containing 12 categories including “background”. The mIOU is compared with other models such as ExfuseNet and DeeplabV3 and it is found that Resnet-Unet achieves a significantly superior performance.

J. Sea-Land Segmentation With Res-UNet And Fully Connected CRF [10]

Z. Chu et al. propose using a UNet segmentation framework and ResNet-18 contraction modules to improve its complex image handling capabilities for a sea-land segmentation task to binarily classify sea and land pixels in remote sensing images of coastlines. They develop their dataset from 208 Google Earth images with 3 bands and since prediction labels from this model are rough, they introduce a fully connected Conditional Random Field (CRF) and a morphological operation to improve their accuracy by about 0.75%. Their results indicate that a UNet framework with ResNet-18 contraction modules could significantly increase prediction accuracy for a lane detection task and suggest that post-processing with a fully connected CRF and morphological operation could reduce noise in the network’s prediction labels.

Problem Statement

The goal of this study is to investigate and compare the performance of two model architectures for lane detection in autonomous driving: the traditional UNet architecture and a modified UNet architecture augmented with a ResNet-50 backbone. The study aims to assess the effectiveness of incorporating ResNet-50, a powerful feature extractor, into the UNet architecture for improved lane detection accuracy by improving global contextual understanding, and mitigating gradient diminishing problems. The evaluation metric is the mean Intersection over Union (mIoU), which measures the pixel-wise overlap between predicted and ground truth lane markings. The experiments will be conducted using the TuSimple Lane dataset of annotated road images and the performance of each model will be analyzed in terms of mIoU scores to determine their efficacy in lane detection tasks.

Dataset

The TuSimple Lane Detection Dataset is a widely-used benchmark dataset designed for lane detection tasks in autonomous driving research [13]. It is a large collection of images captured from real-world driving scenarios, covering diverse road, lighting, and weather conditions. Each image is annotated with pixel-level annotations for lane markings, including lane boundaries, and lane dividers. There are a total of 14,336 lane boundary annotations for 7 different classes such as “Single Dashed”, “Double Dashed” or “Single White Continuous”.This is an extension of the larger TuSimple dataset which consists of road images on US highways with a resolution of 1280x720. Preliminary literature survey indicates its popularity and utility in lane detection research.

To convert the raw data to a usable format for training a model, the data pre-processing involves frame extraction and lane mask creation. For the first step, the last frame of each video clip is extracted, capitalizing on the temporal progression of the footage to potentially capture the most informative scene. Each clip and its frame identifier are then saved together for easier association. For mask creation, provided lane annotations and the path to the raw image file are used and a binary mask image is generated where lanes are delineated. For each lane annotation, non-lane points (indicated by -2 in the annotations) are filtered out and valid lane markers are used to draw white polylines on a black background (see fig.1 and fig.2). The masks are paired with their frames and saved. [16]

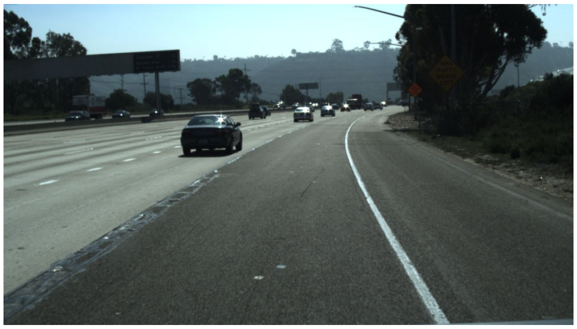


Fig. 1. Ground truth image which is the last frame of video clip. [2]

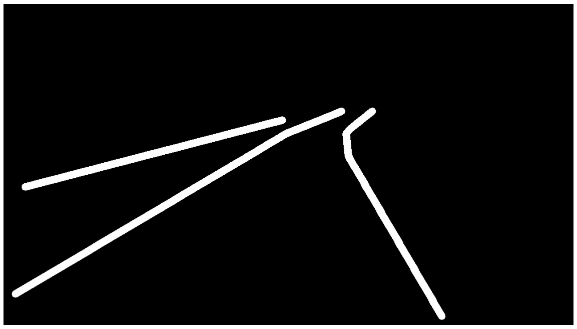


Fig. 2. Binary mask for ground truth image generated through pre-processing.

System Model

This paper proposes an implementation of the two neural network models: UNet and ResUnet which is UNet with a ResNet-50 backbone.

K. UNet based solution [17]

1) Encoder: The encoder has 8 convolution layers, divided into blocks of two with each block followed by a max pooling operation. The blocks capture the image context which helps the model develop an understanding of the scene. Deeper into the encoder, max pooling decreases the spatial resolution, but the feature representation becomes more complex and abstract, allowing the model to identify features such as lane curvature and continuity.

2) Decoder: Then, there is a transition between the encoder and decoder. This bridge has 2 convolution layers followed by dropout.Then each subsequent level of the decoder has 2 convolution layers after the upsampling and concatenation steps, totaling 8 convolution layers in the decoder section, excluding the final 1x1 convolution layer used for output. The decoder gradually restores the spatial dimensions of the encoded features through upsampling and convolution operations. Each step of the decoder combines features from the corresponding encoder level (via skip connections) with the upsampled features, allowing the model to precisely localize and delineate the lanes in the image.

3) Output Layer: The final layer of the model is a convolution layer with a sigmoid activation function. It maps the complex features learned by the network to the probability of each pixel belonging to a lane. The output is a binary segmentation map where each pixel's value indicates the presence or absence of a lane. Thus, we have a total of 19 convolution layers

The model is then compiled with the Adam optimizer and Binary Focal Cross Entropy loss function. This loss is particularly useful for lane detection as the dataset is imbalanced i.e., non-lane pixels vastly outnumber lane pixels.

L. Res-UNet based solution

1) Background: ResNet50 is typically separated into 5 stages, which can be seen below in figure 3. The convolution blocks shown in light blue are slightly different from the ones used in UNet, using 3 convolution layers instead of 2. The main difference comes from the identity block or residual connection, which allows the network to overcome the vanishing gradient problem by adding the stage's original input back to its output after the convolution.

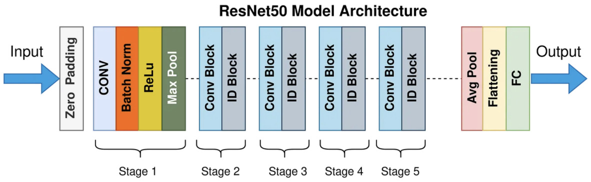


Fig. 3. High-Level ResNet50 Architecture [19]

2) Encoder: The Res-UNet model uses the same decoder architecture as UNet, while the encoding blocks and bridge are replaced by ResNet50 layers, pre-trained on ImageNet and then fine-tuned on the input data. A similar architecture was used by A. K. Hernanda et al. with ResNet32, shown below in figure 4.

Just like with the original UNet architecture, our model had 4 encoding blocks, a bridge, and 4 decoders to recover the images original resolution. The input to each encoding block consists of the output from the decoding block (or bridge) below, as well as the output from the encoding block at the same level.

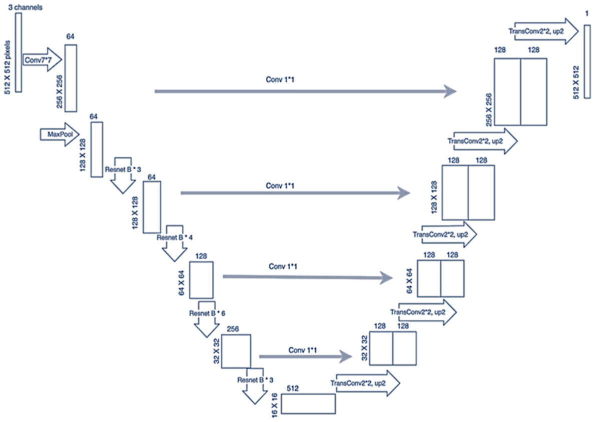


Fig. 4. UNet model with ResNet encoder [2]

The first encoding block was replaced with the input to ResNet50, the second was the output after performing convolution with 7x7 kernel size and stride of 2 [18]. The third encoding block came from the second convolution layer in block 3 of ResNet50, while the fourth came from the third convolution layer of block 4 [18]. The bridge also came from ResNet50, being the fourth convolution layer in the 6th block of ResNet50. The full architecture diagram can be seen below in figures 5 and 6.

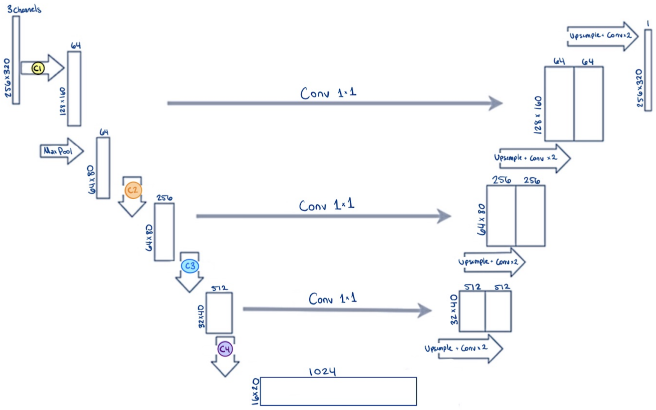


Fig. 5. UNet model with ResNet50 encoder

Initially, pre-trained ResNet50 on ImageNet was being used, but it was found that pre-training reduced performance, with especially worse IoU scores. Visually, this resulted in many blotchy and disconnected patches in the prediction masks.

One additional change made to the architecture was the use of a Leaky ReLU activation in the decoding arm, instead of a normal ReLU activation. Leaky ReLU has been shown to help reduce the dying neurons issue, which is when many of the neurons become deactivated (outputting zero) because it has negative inputs. This was found to improve performance likely due to more neurons remaining active.

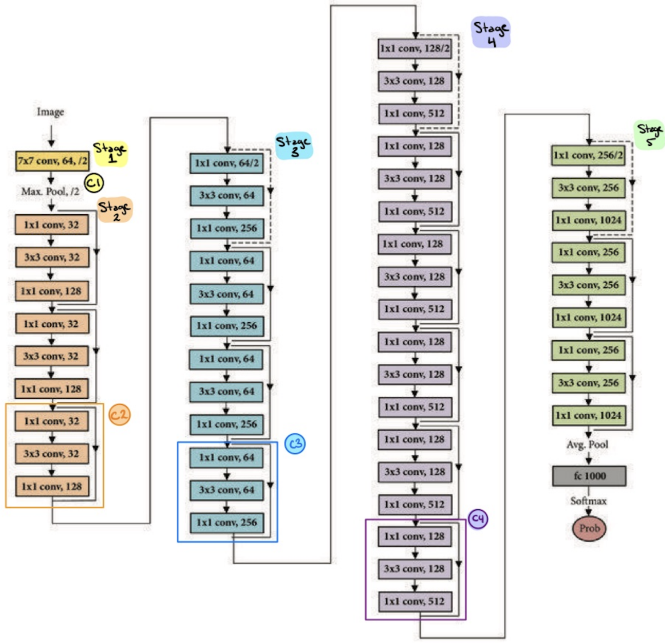


Fig. 6. ResNet50 with labeled blocks used in Res-UNet [20]

RESULTS

M. Training

The dataset consisted of 2000 total samples, with 200 being used for validation and 200 for testing, leaving 1600 samples for training. The models were compiled with the Adam optimizer and Binary Focal Cross Entropy loss function. This loss is particularly useful for lane detection as the dataset is imbalanced with non-lane pixels vastly outnumbering lane pixels.

Training was done on the V100 GPU with High RAM in Google Colab, which consumed approximately 5 compute units per hour. UNet had a total of 10 million more trainable parameters than Res-UNet, resulting in training time per epoch taking approximately 46 seconds compared to 34 seconds.

Through experimentation it was found that UNet performed best when trained with 15 epochs, while Res-UNet performed best with 10. The accuracy and loss training curves for UNet and Res-UNet can be seen below in figure 7 and 8 respectively.

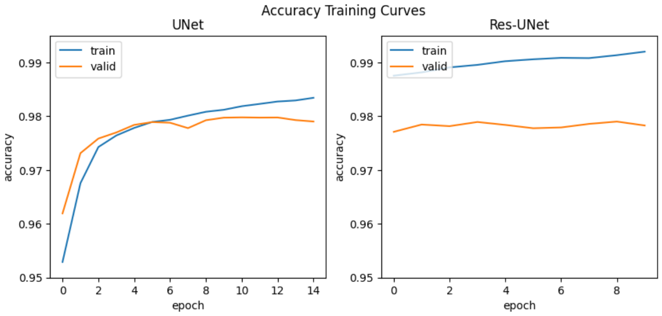


Fig. 7. Accuracy Training Curves for UNet (left) and Res-UNet (right)

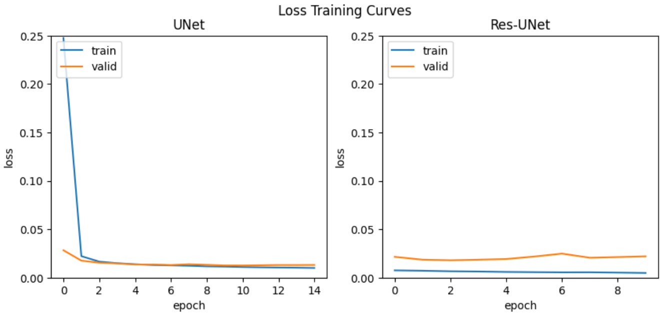
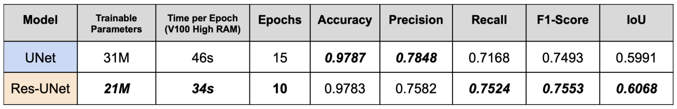


Fig. 8. Loss Training Curves for UNet (left) and Res-UNet (right)

N. Performance

The speed and performance of UNet and Res-UNet can be seen summarized below in table A, with the better scores in bold. Res-UNet outperformed UNet for all three speed sections, having 10M fewer trainable parameters, <10 second shorter training time per epoch on the V100 High RAM GPU, and required 5 fewer epochs to finish training. Both models were trained on 5, 10, and 15 epochs and the final scores listed below correspond to the best trial, which was 15 epochs for UNet and 10 epochs for Res-UNet.



Model	Trainable Parameters	Time per Epoch (V100 High RAM)	Epochs	Accuracy	Precision	Recall	F1-Score	IoU
UNet	31M	46s	15	***0.9787***	***0.7848***	0.7168	0.7493	0.5991
Res-UNet	***21M***	***34s***	**10**	0.9783	0.7582	***0.7524***	***0.7553***	***0.6068***

Fig. 9. Performance scores of UNet and ResUNet on TuSimple dataset

UNet reported higher scores in accuracy and precision, however the accuracy scores for both models are essentially the same. Res-UNet had higher scores in recall, F1-score, and IoU. A visual comparison of these results can be seen in figure 10 below.

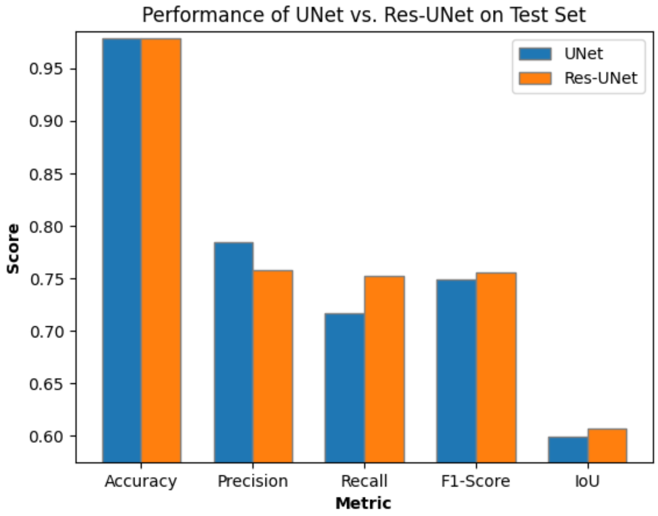


Fig. 10. Performance scores on testing set with UNet vs. Res-UNet

Conclusion

In conclusion, the purpose of this project was to compare the performance of UNet against a UNet and ResNet model on the task of lane detection. The UNet model was chosen as previous works indicated its ability to perform well in lane detection tasks. The other model, UNet combined with ResNet, was selected after studies showed that the combination is robust in other similar image segmentation tasks such as in the biomedical field. Thus, the two models were compared using the TuSimple lane detection dataset using evaluation metrics such as mIoU. It was discovered that the Res-UNet model had higher scores in recall, F1 and IoU while UNet had higher accuracy and precision scores. Next steps include comparing UNet to UNet and ResNet using a more complex lane detection dataset. This is because the Res-UNet architecture has better capabilities to understand more complex scenes and may outperform UNet noticeably when trained and tested on a more challenging dataset. This will allow further exploration of the models' effectiveness. In addition, another next step can also be post-processing the results to mitigate the effects of objects on the lane markings in the original image.

References

[1] L. -A. Tran and M. -H. Le, ”Robust U-Net-based Road Lane Markings Detection for Autonomous Driving,” 2019 International Conference on System Science and Engineering (ICSSE), Dong Hoi, Vietnam, 2019.

[2] R. SaiNikhil, S. G. Rao and P. V. P. Rao, ”Identification of Road and Surrounding Obstacles using U-Net Architecture for Better Perception,” 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2023.

[3] Y. Luo, Y. Zhang and Z. Wang, ”Lane line detection algorithm based on improved UNet network,” 2023 8th International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), Okinawa, Japan, 2023.

[4] JR. K. Kaushal, A. R, K. K. B. Giri, M. Sindhu, N. L and B. Ronald, ”Deep Learning Based Segmentation Approach for Automatic Lane Detection in Autonomous Vehicle,” 2023 International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS), Erode, India, 2023.

[5] S. -H. Mousavi and K. -C. Yow, ”Self-Attention blocks in UNet and FCN for accurate semantic segmentation of difficult object classes in autonomous driving,” 2023 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), Regina, SK, Canada, 2023.

[6] E. Brahim, E. Amri and W. Barhoumi, ”Enhancing Change Detection in Spectral Images: Integration of UNet and ResNet Classifiers,” 2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI), Atlanta, GA, USA, 2023.

[7] A. K. Hernanda, I. K. Eddy Purnama, E. Mulyanto Yuniarno and J. Nugroho, ”Semantic Segmentation of Venous on Deep Vein Thrombosis (DVT) Case using UNet-ResNet,” 2022 10th International Conference on Information and Communication Technology (ICoICT), Bandung, Indonesia, 2022.

[8] T. A. Youssef, B. Aissam, D. Khalid, B. Imane and J. E. Miloud, ”Classification of chest pneumonia from x-ray images using new architecture based on ResNet,” 2020 IEEE 2nd International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS), Kenitra, Morocco, 2020.

[9] Z. Li, W. Liu, G. Wu and S. Yang, ”Semantic segmentation of unmanned aerial vehicle image based on Resnet-Unet,” 2023 8th International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), Okinawa, Japan, 2023.

[10] Z. Chu, T. Tian, R. Feng and L. Wang, ”Sea-Land Segmentation With Res-UNet And Fully Connected CRF,” IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 2019.

[11] D. R. Soumya, D. L. K. Reddy, A. Nagar and A. K. Rajpoot, ”Enhancing Brain Tumor Diagnosis: Utilizing ResNet-101 on MRI Images for Detection,” 2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN), Vellore, India, 2023.

[12] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," PMLR, https://proceedings.mlr.press/v78/dosovitskiy17a.html (accessed Feb. 25, 2024).

[13] "Autonomous Trucks: Self-driving truck: Driverless vehicles," TuSimple, https://www.tusimple.com/ (accessed Feb. 25, 2024).

[14] Organisation for Economic Co-operation and Development, "Catalogue of Tools & Metrics for Trustworthy AI," OECD.AI, https://oecd.ai/en/catalogue/metrics/mean-intersection-over-union-%28iou%29 (accessed Feb. 25, 2024).

[15] Hikmatullahmohammadi, "Road lane detection with UNET- tusimple dataset," Kaggle, https://www.kaggle.com/code/hikmatullahmohammadi/road-lane-detection-with-unet-tusimple-dataset (accessed Apr. 12, 2024).

[16] Hikmatullahmohammadi, "Road Lane Line [tusimple] dataset preparation," Kaggle, https://www.kaggle.com/code/hikmatullahmohammadi/road-lane-line-tusimple-dataset-preparation (accessed Apr. 12, 2024).

[17] S. Mukherjee, "The annotated resnet-50," Medium, https://towardsdatascience.com/the-annotated-resnet-50-a6c536034758 (accessed Apr. 12, 2024).

[18] K. He, "Deep Residual Learning for Image Recognition," The Computer Vision Foundation, https://www.cv-foundation.org/openaccess/contentcvpr2016/papers/ HeDeepResidualLearningCVPR2016paper.pdf (accessed Apr. 12, 2024).

[19] S. Mukherjee, "The annotated resnet-50," Medium, https://towardsdatascience.com/the-annotated-resnet-50-a6c536034758 (accessed Apr. 12, 2024).

[20] A. Qamar Bhatti et al., "Explicit content detection system: An approach towards a safe and ethical environment," Applied Computational Intelligence and Soft Computing, https://www.hindawi.com/journals/acisc/2018/1463546/ (accessed Apr. 12, 2024).