A Comparative Study of U-Net and DeepLabV3 for Autonomous Vehicle Perception

1st Kenneth McAlinden
Faculty of Science
University of Western Ontario
London, Canada
kmcalin@uwo.ca

2nd Piranavan Jeyakumar
Faculty of Science
University of Western Ontario
London, Canada
pjeyaku2@uwo.ca

3rd Tudor Stoica
Faculty of Science
University of Western Ontario
London, Canada
tstoica2@uwo.ca

4th Yuvraj Virdi
Faculty of Science
University of Western Ontario
London, Canada
yvirdi@uwo.cs

***Abstract*—We delve into a comparative study between DeepLabV3 and U-Net models for semantic image segmentation tailored for autonomous vehicles. Semantic image segmentation is pivotal for autonomous vehicles as it enables the detection of other vehicles, pedestrians, and the surrounding environment. DeepLabV3, and U-Net, are both prominent convolutional neural network architectures for semantic image segmentation. There remains a research gap concerning the computational efficiency of these networks and comprehensive comparative studies between different segmentation architectures.**

Our research objectives revolve around setting up and training both models using the KITTI dataset, performing hyperparameter optimization, and evaluating their performance metrics such as inference time, model pixel accuracy, average recall, average precision, and average F1-score. Our results highlight the effectiveness of DeepLabV3 in capturing intricate features and its competitive performance metrics, showcasing its potential for real-world deployment in autonomous vehicle perception tasks.

The impact of our findings on theory and practice is noteworthy. The analysis of performance disparities between DeepLabV3 and U-Net enhances theoretical frameworks related to feature extraction, context aggregation, and spatial reasoning in semantic segmentation. Moreover, our results offer practical guidance for selecting suitable segmentation models for autonomous driving applications, facilitating informed decision-making based on accuracy, computational efficiency, and real-time performance. This understanding is necessary for optimizing autonomous vehicle performance, enhancing scene identification, object detection, and decision-making processes.

Moving forward, future research endeavors could focus on optimizing both models to improve their inference speed without compromising accuracy. Additionally, expanding the dataset and exploring other variants of deep learning-based segmentation models could provide further insights and advancements in semantic segmentation for autonomous driving scenarios.

The report begins with an introduction that provides context, background, and research gaps in the field of semantic image segmentation for autonomous vehicles. This is followed by a detailed explanation of research objectives and the type of work conducted, leading into a comprehensive presentation of the results obtained from the comparative study between DeepLabV3 and U-Net. The report then highlights the novelty and progress made, along with an analysis of the impact of the results on both theory and practice. Finally, the conclusions section summarizes the key findings and suggests avenues for future research, ensuring a coherent and informative structure throughout the document.

I. INTRODUCTION

A. Context

In this report we will be comparing the DeepLabv3 [27] and U-Net [28] models on semantic image segmentation for autonomous vehicles. DeepLabV3 is a modern convolutional neural network (CNN) for semantic image segmentation, introduced by Google in 2017 [13]. DeepLabV3 incorporates several key components, including atrous convolution to capture multi-scale context, atrous spatial pyramid pooling (ASPP) to aggregate information at multiple scales, and conditional random fields (CRFs) for post-processing to refine boundaries. U-Net, which is a widely adopted CNN for biomedical image segmentation, was proposed by Ronneberger et al. in 2015 [14]. U-Net is particularly well-suited for biomedical image segmentation tasks due to its ability to handle small training datasets effectively and produce accurate pixel-level predictions.

B. Background & Related Work

The first breakthrough in this domain began with the introduction of fully convolutional networks (FCNs) [9]. FCNs presented an approach to pixel-wise classification and leveraging transposed convolution for upsampling. FCNs incorporated a skip architecture to refine segmentation output by utilizing higher-resolution feature maps. This method laid the foundation for subsequent advancements in segmentation accuracy. Following this, various strategies such as multi-scale approaches, structured models, and spatio-temporal architectures were introduced, each aiming to enhance segmentation accuracy. All of these methodologies prioritized the refinement of segmentation results for improved accuracy and robustness. Additionally, the advent of well-established benchmarks and datasets like Cityscapes [10] and Mapillary [11] incited competition, driving further enhancements in accuracy across the field.

C. Research Gaps

One major gap in the research of this field is the computational efficiency of these networks and comparative studies between the different segmentation architectures. Autonomous vehicles rely on semantic image segmentation for the detection of other vehicles, pedestrians, and the surrounding

road environment. Meaning that the networks employed by the vehicle must process all these factors within milliseconds while driving on the road. As such, it would be crucial to have a detailed comparison of all of the different networks and architectures [12] to determine which would be the most optimal (based on speed, accuracy, resource load, etc) to employ in real world applications.

D. Research Objectives and Type of Work Done

The following objectives are applied to BOTH of our models, DeepLabv3 and U-Net:

1) ID-1: Set up untrained models and adjust them to be trained on the KITTI [26] dataset.
2) ID-2: Train both models using the KITTI dataset (dataset was split into training and validation sets to measure training/testing error).
3) ID-3: Perform hyperparameter optimization on both models to identify the optimal hyperparameter configuration for both models on the KITTI dataset.
4) ID-4: Use the optimal hyperparameter configuration to define a final version of both models and train these final versions on the KITTI dataset.
5) ID-5: Perform a comparative study by measuring the speed, accuracy, and performance of both final models. These results will be used determine which of the two models is generally better for application in autonomous driving semantic segmentation.

E. Results Obtained

In this section, we present the findings of our comparative study on DeepLabV3 and U-Net for autonomous driving semantic segmentation. The evaluation encompasses various aspects crucial for model performance evaluation.

We begin by comparing the visualization results obtained from the DeepLabV3 and U-Net models on the validation set and delving into an analysis of the discrepancies observed in the predicted masks generated by both models.

We then discuss key performance metrics essential for evaluating segmentation models tailored to autonomous driving scenarios. Inference time, denoting the duration for processing input data and generating predictions, serves as a vital indicator of model efficiency. Model pixel accuracy, a fundamental metric, quantifies the accuracy of pixel-wise predictions. Additionally, we delve into average recall, average precision, and average F1-score metrics. These metrics provide insights into the model's performance across various classes within the KITTI dataset.

These quantitative evaluations lay the groundwork for a comprehensive understanding of the performance discrepancies between DeepLabV3 and U-Net in the context of autonomous driving semantic segmentation.

F. Novelty/Progress Made

The key novelty of our custom DeepLabV3 and U-Net implementations lies in their adoption of the ResNet-50 backbone, which enables the model to capture intricate features in the input images effectively. Additionally, DeepLabV3's competitive performance metrics demonstrate its suitability for autonomous vehicle perception tasks, showcasing its potential for real-world deployment, given the small dataset it was trained on. While both models are commonly used for semantic segmentation tasks, comparing them specifically for autonomous driving applications is less common in literature. Evaluation of these models through the mentioned performance metrics provides a comprehensive understanding of both models' performance within this domain.

G. Impact of Results on Theory and Practice

The comparative study sheds light on the performance disparities between DeepLabV3 and U-Net in the context of autonomous driving semantic segmentation. The analysis of visualization results and performance metrics provides insights into the underlying mechanisms driving the segmentation decisions of DeepLabV3 and U-Net, enhancing key aspects of theoretical frameworks such as feature extraction, context aggregation, and spatial reasoning.

The findings offer practical guidance for selecting suitable segmentation models for autonomous driving applications, enabling decision-makers to make informed choices based on speed, accuracy, and real-time performance. This understanding also facilitates the optimization of autonomous vehicle performance, as insights into the impact of different segmentation models on perception tasks enhance scene identification, object detection, and decision-making.

II. RESEARCH METHODOLOGY

A. Model Selection

Due to their prominent use in image segmentation tasks [17] [18] and contrasting architectural designs we used DeepLabv3 and U-Net models. These choices offer us robust comparisons for performance, specifically in autonomous driving contexts. Both use ResNet as a backbone, a common and effective choice for image segmentation [19]. We specifically used ResNet50 because it is the most intermediate in terms of depth and computational complexity and we ensured each model was initialized with no encoded weights to not skew our data.

B. Data Collection

We decided to use the KITTI dataset, an industry standard for bench-marking in autonomous driving research [20]. This dataset is relatively small, consisting of only 200 annotated images with a wide assortment of environments. We selected a smaller dataset because of our limited computing resources and its ability to provide a better understanding of the models learning efficiency.

C. Data Preparation

Our main focus in data preparation was to ensure the data would be optimized for image segmentation deep learning according to industry standards. The following three preprocessing steps were taken:

1) Implementation of random horizontal flips, known to help models generalize through adding variations to the dataset [21]. We opted against using vertical flips or image rotations as they introduce unrealistic scenarios, such as images displaying roads above the sky [22].
2) We applied normalization using the calculated mean and standard deviation of our dataset. Normalization was chosen as it is a very common data pre-processing step in the field of computer vision [21].

D. Common Training Regimen

An identical training regimen was used for both models to guarantee a direct comparison. The factors we ensured to be constant were dataset processing and optimization algorithms. To ensure the processing towards optimization was equal we used the same set of hyper-parameters and number of trials. This way a model can use its optimal parameters given a set amount of processing, allowing for a comparison between equal, optimal models.

- Hyper-parameter Tuning: We used Optuna [23] for hyper-parameter optimization, optimizing the learning rate and batch size of the training. Optuna was chosen for its ability to provide plots based on its results. The set of hyper-parameters was chosen using hyper-parameter importance plots provided by Optuna, shown in Figure 1.

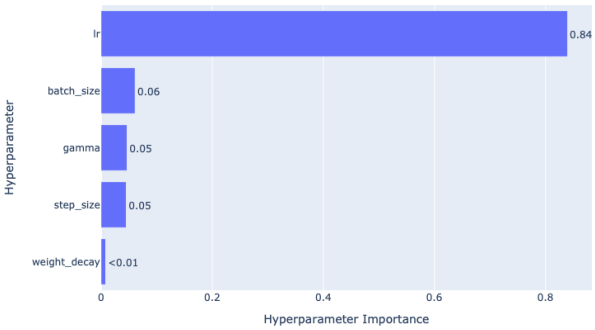

Fig. 1. Hyper-parameter Importance

Due to computational constraints, only the two most important hyper-parameters, namely learning rate and batch size, were chosen and optimized for their respective models. Provided is the resulting parallel coordinate, shown in Figure 2 and Figure 3.

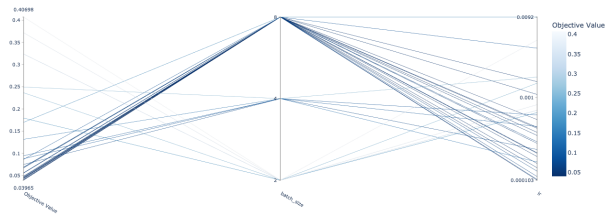

Fig. 2. Hyper-parameter Importance for DeepLabv3

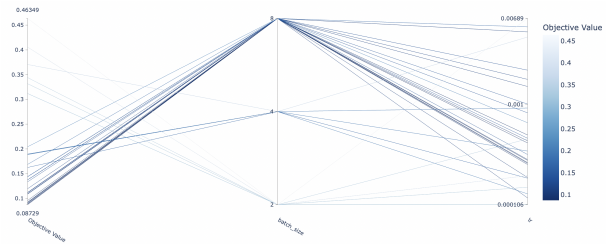

Fig. 3. Hyper-parameter Importance for U-Net

- Training Strategy: To gauge model performance we used a common 80:20 train-validation split, rather than 70:30, as it was crucial to provide the model enough training data to reliably detect features. After, the optimal hyper-parameters for each model were used to train the final model using an AdamW optimizer to minimize Cross Entropy Loss. The AdamW optimizer was selected because it performed best with our models when compared to other common PyTorch optimizers, such as Stochastic Gradient Descents. The optimization criteria for Cross Entropy Loss was selected for its prevalent use in classification tasks [23].

E. Performance Evaluation

We evaluated the model using various metrics, including Intersection over Union (IoU), Precision, Recall, and F1-Score, both at the class and category levels. These metrics are industry standards for bench-marking image segmentation models [25]. Using them allows us to accurately compare each model's advantages and shortcomings when performing segmentation tasks.

F. Threats to Validity

Our main concern was ensuring consistency between models because we wanted any observable differences to be attributed to the model architecture and not the cause of extraneous variables. Such variables include data pre-processing, augmentation techniques, and optimization.

III. RESULTS

In this section, we delve into testing and evaluation results of the deep learning segmentation models tailored towards autonomous driving semantic segmentation. The discussion is structured into two sections, each addressing different aspects of the evaluation process. The first section provides an overview about the visualization results of the segmentation models. The second section offers performance metrics such as Inference Time, Model Pixel Accuracy, Intersection over Union (IoU), Average Precision, Average Recall, and Average F1-Score.

A. Visualization Results of Segmentation Models for Autonomous Driving

The KITTI dataset was used for training and validation of the models, containing diverse images of urban streets,

buildings, pedestrians, vehicles, and other urban objects. From Figure 4 and Figure 5, the models efficiently detect and segment the various classes within the images.

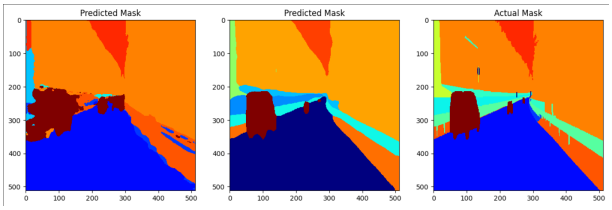

Fig. 4. Comparison between U-Net and DeepLabV3's predicted mask and the actual mask of image 000191_10.png of the training set in KITTI

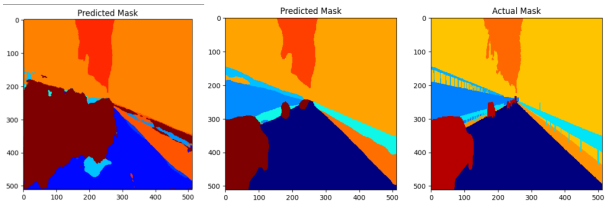

Fig. 5. Comparison between U-Net and DeepLabV3's predicted mask and the actual mask of image 000194_10.png of the training set in KITTI

From Figure 4 and Figure 5 it is clear that both U-Net and DeepLabV3 can correctly segment the road (as dark blue), the sky (as orange-red), and the surrounding trees (as orange). Where DeepLabV3 shines (and U-Net falls short) is the proper segmentation of the finer details, such as the outlines of the cars and the roadside barriers. U-Net's short-comings for the semantic segmentation of the KITTI dataset images are a result of its shallow architecture, resulting in an insufficient capacity to handle the complexity of the segmentation task, and its inability to capture extensive contextual information across different scales

B. Performance Metrics

Various matrices exist to assess and gauge the precision of segmentation methods [15], the ones used in this work are as follows:

1) Inference Time: Inference time refers to the duration it takes for the deep learning segmentation model to process input data and generate predictions. It measures the efficiency of the model's prediction process, which influences its use in autonomous driving where rapid decision-making is crucial. Both model's inference times were compared using a NVIDIA RTX A1000 GPU.

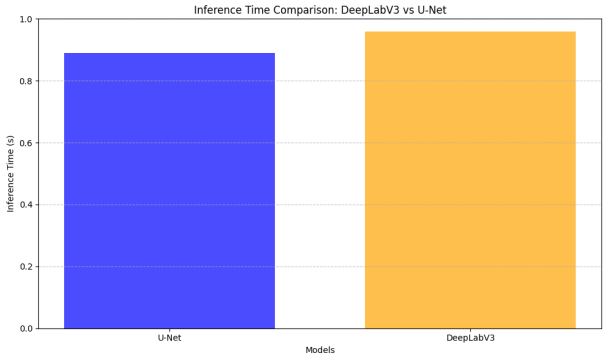

Fig. 6. Inference Time comparison between U-Net and DeepLabV3 semantic segmentation models.

2) Model Pixel Accuracy: Pixel Accuracy is one of the most widely used metrics of evaluation for segmentation models. As the name suggests, it is the accuracy of the model's pixel-wise prediction:

$$P_{accuracy} = \frac{\sum_{i=0}^{N} p_{ii}}{\sum_{i=0}^{N} \sum_{j=0}^{N} P_{ij}} \tag{1}$$

For the above formula, N is the total number of pixels in the validation image, p_{ij} are the pixels of the true mask (the number of pixels of class i predicted as class j), and p_{ii} is the predicted pixels as class i (predicted mask).

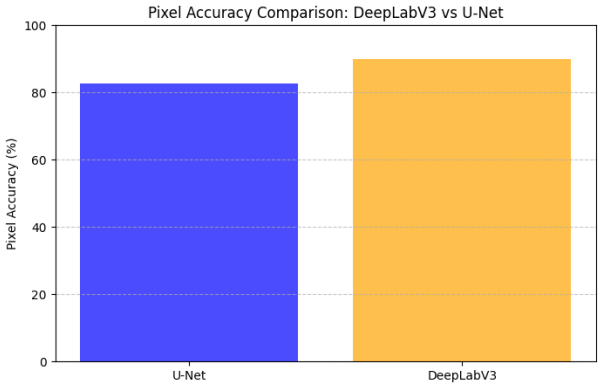

Fig. 7. Pixel Accuracy comparison between U-Net and DeepLabV3 semantic segmentation models.

3) Average Recall, Average Precision, and Average F1-Score: Recall, Precision, and F1-Score are popular metrics of evaluation for segmentation models as well. We calculate the Recall, Precision, and F1-Score of each class in the KITTI dataset and calculate the average to determine the model's general performance on the dataset. Recall, Precision, and F1-Score for each class is defined as:

$$Recall = \frac{tp}{tp + fn} \tag{2}$$

$$Precision = \frac{tp}{tp + fp} \tag{3}$$

For the above formulas, each variable corresponds to a segment of the confusion matrix, where tp is the True Positive, fn is the False Negative, and fp is the False Positive.

$$F1 - Score = \frac{(2 \times Precision) \times Recall}{Precision + Recall} \tag{4}$$

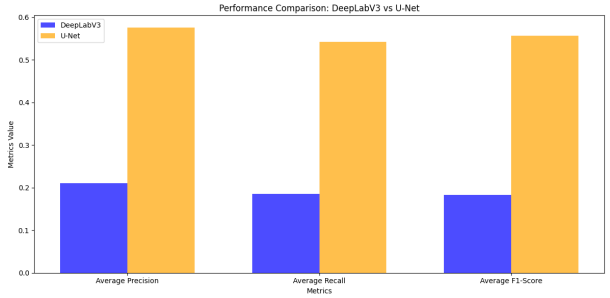

Fig. 8. Comparison of Average Precision, Average Recall, and Average F1-Score of all the classes from the KITTI dataset between U-Net and DeepLabV3 semantic segmentation models.

4) Intersection-over-Union (IoU): Intersection-over-Union is the ratio of intersection and union area between the predicted segmentation mask and the true mask, defined as:

$$IoU = J(A, B) = |A \cap B|/|A \cup B| \tag{5}$$

For the above formula, A represents the true mask of the image and B is the predicted segmentation mask. Given that the KITTI dataset deals with 34 unique classes which the image could be segmented into, the scores of the top 8 most prominent classes, Vehicle, Human, Sky, Nature, Object, Construction, Flat, and Void will be compared for both models.

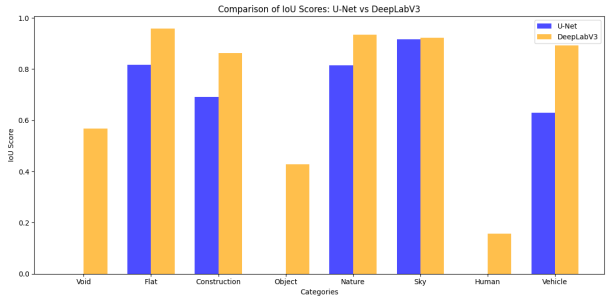

Fig. 9. Comparison of Intersection-over-Union scores of top 8 prominent classes between U-Net and DeepLabV3 semantic segmentation models.

The overall performance results of our DeepLabV3 and U-Net models depicted in Figures 1 - 5 indicate that the DeepLabV3 model does a better job segmenting and classifying the segments correctly compared to the U-Net model. This difference in performance between models on the KITTI dataset, which contains scenes with intricate details and precise object boundaries, highlights DeepLabV3's ability to capture fine details effectively and accurately delineating object boundaries.

IV. CONCLUSIONS & FUTURE WORK

In this comparative study, we investigated the performance of two deep-learning semantic segmentation models, DeepLabV3 and U-Net, in the context of autonomous driving semantic segmentation. Our custom DeepLabV3 model delivered promising results, achieving a high accuracy of 90% on the validation dataset. However, while the model demonstrated a moderate performance in terms of Average Precision (0.58), Average Recall (0.54), and Average F1-Score (0.56), its Inference Time of 0.96 seconds suggests room for optimization to enhance real-time performance.

On the other hand, our custom U-Net model also demonstrated strong accuracy, reaching 82.7% on the validation set. However, its Average Precision (0.2105), Average Recall (0.1857), F1-Score (0.1831), and Inference Time of 0.89s suggest areas for improvement and optimization. It's important to note that these metrics provide insights into the model's ability to accurately classify objects in the autonomous driving environment, with higher scores indicating better performance and lower Inference Time indicating faster speeds.

Furthermore, reducing Inference Time is crucial for real-time applications such as autonomous driving, where timely decision-making is paramount. Therefore, future research efforts could focus on optimizing both models to improve their Inference Time without compromising accuracy.

In addition, expanding the dataset to include the entirety of the CityScapes dataset [16] could provide a more comprehensive evaluation of the models' performance across various scenarios. Exploring other variants of deep learning-based segmentation models for comparison could also offer valuable insights into alternative approaches for semantic segmentation in autonomous driving scenarios.

REFERENCES

[1] "Computer vision challenges in autonomous vehicles: The future of AI", 2023. https://www.superannotate.com/blog/computer-vision-in-autonomous-vehicles

[2] Nilesh Barla, "Self-Driving Cars With Convolutional Neural Networks (CNN)", 2023. https://neptune.ai/blog/self-driving-cars-with-convolutional-neural-networks-cnn

[3] Hiten Patel, "Image Classification For Autonomous Vehicles", 2020. https://github.com/hpatel530/Image-Classification-for-Autonomous-Vehicles-

[4] Hironobu Fujiyoshi, Tsubasa Hirakawa, Takayoshi Yamashita, "Deep learning-based image recognition for autonomous driving", 2019. https://www.sciencedirect.com/science/article/pii/S0386111219301566

[5] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, "Going Deeper with Convolutions", 2014.

[6] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, Kaiming He, "Real-Time Flying Object Detection with YOLOv8", 2023.

[7] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, Liang-Chieh Chen, "MobileNetV2: Inverted Residuals, Linear Bottlenecks", 2019.

[8] Mingxing Tan, Ruoming Pang, Quoc V. Le, "EfficientDet: Scalable and Efficient Object Detection", 2020.

[9] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 640-651, April 2017.

[10] M. Cordts et al., "The Cityscapes Dataset for Semantic Urban Scene Understanding," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 99, pp. 1-1, 2024.

[11] G. Neuhold et al., “The Mapillary Vistas Dataset for Semantic Understanding,“ in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4998-5007.
[12] I. Ahmed et al., “Comparison of Deep-Learning-Based Segmentation Models: Using Top View Person Images,“ in IEEE Access, vol. 2, pp. 1107-1125, 2014.
[13] L.-C. Chen et al., “Rethinking Atrous Convolution for Semantic Image Segmentation,“ arXiv:1706.05587, 2017.
[14] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,“ in Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015, pp. 234-241.
[15] S. Minaee et al., “Image Segmentation using Deep Learning: A Survey,” IEEE Xplore, 2022.
[16] Cityscapes Dataset, “CityScapes Dataset Overview,“ [Online]. Available: https://www.cityscapes-dataset.com/. Accessed on: Apr. 4, 2024.
[17] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 4, pp. 834–848, Apr. 2018, doi: https://doi.org/10.1109/tpami.2017.2699184.
[18] M. Siam, M. Gamal, M. Abdel-Razek, S. Yogamani, M. Jagersand and H. Zhang, ”A Comparative Study of Real-Time Semantic Segmentation for Autonomous Driving,” 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 2018, pp. 700-70010, doi: 10.1109/CVPRW.2018.00101.
[19] M. B. Sahaai, G. R. Jothilakshmi, D. Ravikumar, R. Prasath, and S. Singh, “ResNet-50 based deep neural network using transfer learning for brain tumor classification,” INTERNATIONAL CONFERENCE ON RECENT INNOVATIONS IN SCIENCE AND TECHNOLOGY (RIST 2021), 2022, doi: https://doi.org/10.1063/5.0082328.
[20] I. Papadeas, L. Tsochatzidis, A. Amanatiadis, and I. Pratikakis, “Real-Time Semantic Image Segmentation with Deep Learning for Autonomous Driving: A Survey,” Applied Sciences, vol. 11, no. 19, p. 8802, Sep. 2021, doi: https://doi.org/10.3390/app11198802.
[21] C. Shorten and T. M. Khoshgoftaar, “A survey on Image Data Augmentation for Deep Learning,” Journal of Big Data, vol. 6, no. 1, Jul. 2019, doi: https://doi.org/10.1186/s40537-019-0197-0.
[22] K. Alomar, H. I. Aysel, and X. Cai, “Data Augmentation in Classification and Segmentation: A Survey and New Strategies,” Journal of Imaging, vol. 9, no. 2, p. 46, Feb. 2023, doi: https://doi.org/10.3390/jimaging9020046.
[23] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, “Optuna,” Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Jul. 2019, doi: https://doi.org/10.1145/3292500.3330701.
[24] [1] J. Tian, N. Chowdhury, H.-P. Chiu, and Z. Seymour, “Striking the right balance: Recall loss for semantic segmentation,” Research Gate, https://www.researchgate.net/publication/353061904_Striking_the_Right_Balance_Recall_Loss_for_Semantic_Segmentation.
[25] D. Müller, I. Soto-Rey, and F. Kramer, “Towards a guideline for evaluation metrics in medical image segmentation,” BMC Research Notes, vol. 15, no. 1, Jun. 2022, doi: https://doi.org/10.1186/s13104-022-06096-y.
[26] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, “Vision meets robotics: The KITTI dataset,” The International Journal of Robotics Research, vol. 32, no. 11, pp. 1231–1237, Aug. 2013, doi: https://doi.org/10.1177/0278364913491297.
[27] P. Iakubovskii, ”Segmentation Models,” GitHub, 2019. [Online]. Available: https://github.com/qubvel/segmentation_models.
[28] TorchVision maintainers and contributors, ”TorchVision: PyTorch’s Computer Vision library,” GitHub, 2016. [Online]. Available: https://github.com/pytorch/vision.