Scene Classification in Rural Environments for Autonomous Driving Using U-Net and ResNet

Victor M. A. do Nascimento[1*], André T. Cunha Lima[2]
[1,2] *SENAI CIMATEC University, MCTI, Salvador, Bahia, Brazil*
[2] *Federal University of Bahia, Institute of Physics, Department of Earth and Environmental Physics, Salvador, Bahia, Brazil*
**Corresponding author: SENAI CIMATEC University, Av. Orlando Gomes, 1845 - Piatã, Salvador - BA, 41650-010*
victor.nascimento@aln.senaicimatec.edu

Abstract: This paper addresses the semantic segmentation of road images, a critical task for autonomous vehicle navigation, particularly in non-urban environments that present significant challenges. While much research focuses on well-maintained roads in developed countries, this study confronts the complexities of real-world conditions, such as those prevalent in developing nations, which feature vast networks of unpaved and poorly maintained roads. The core of our methodology is a neural network architecture that synergistically combines the encoder-decoder structure of U-Net with the feature extraction power of a ResNet backbone. The primary objective is the precise classification of each image pixel into one of four essential categories for navigation: background, asphalt, paved, and unpaved road. The model's training regimen involved exploring different ResNet versions (ResNet18, ResNet34, and ResNet50) as the encoder backbone to assess the impact of network depth. A key aspect of our approach was a progressive training strategy, where model versions were trained on images of varying resolutions. The results demonstrated a significant and somewhat counter-intuitive finding: training the ResNet34-U-Net model with images at half the original resolution yielded the best overall performance, achieving the highest Dice and IoU scores. This suggests that reducing image resolution acts as an effective form of regularization, compelling the model to learn more general and robust features by ignoring minor, irrelevant details. This outcome not only enhances the model's generalization capabilities for diverse and imperfect road conditions but also carries a substantial practical advantage by reducing the computational cost of training and inference, a crucial factor for deployment on resource-constrained embedded systems in autonomous vehicles.

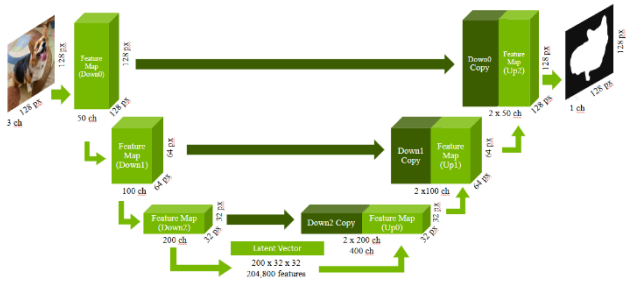
Keywords: Semantic Segmentation, Computer Vision, Neural Networks, Rural Roads

1. Introduction

Detecting a navigable path is an important function in visual navigation systems, where several challenges must be considered, such as terrain variations, lighting changes, and the presence of potholes or water puddles. Data from the National Transport Confederation of 2022 [1] show that only 15.81% of the Brazilian road network consists of paved roads, highlighting the challenge for the widespread adoption of autonomous vehicles in regions distant from large urban centers. Studies of this kind often utilize road scenarios from developed countries, which typically exhibit little to no variation in terrain surface [2]. Some works involve systematic literature reviews: for instance, one 11-year review analyzed computer vision methods applied to this type of problem, emphasizing the surface types these methods address, their adaptability to surface changes, and their capability to distinguish potential road defects or changes, such as potholes, shadows, and water puddles [3]. Convolutional Neural Networks (CNNs), the U-Net architecture, and Residual Networks (ResNet) have emerged as prominent deep learning architectures employed in this context. Research has primarily focused on improving semantic segmentation accuracy, acknowledging the ongoing challenges in achieving reliable environmental perception under diverse and complex driving conditions [4]. Convolutional Neural Networks (CNNs) represent a class of deep learning models whose

fundamental architecture comprises convolutional layers, pooling layers, and activation functions. Convolutional layers act as feature extractors, learning hierarchical spatial patterns in the input data through the application of convolutional filters [4]. Pooling layers are responsible for reducing the spatial dimensionality of feature maps, providing translational invariance and decreasing sensitivity to minor input perturbations. Activation functions, such as ReLU (Rectified Linear Unit), introduce nonlinearity into the model, enabling it to learn complex relationships within the data [5]. As show in Figure 1, the U-Net is a convolutional neural network architecture characterized by its encoder-decoder structure.

Figure 1. U-Net network architecture [23]



The architecture consists of a contracting path (encoder) and an expanding path (decoder). The contracting path follows the typical CNN architecture, comprising multiple convolutional layers followed by max-pooling layers. These progressively reduce the spatial resolution of the feature maps while increasing the number of feature channels. This process allows the encoder to capture high-level contextual information from the input image [5]. Residual Networks (ResNets) address the vanishing gradient problem that typically hinders the training of very deep networks [6]. The main contribution of ResNet is the introduction of residual blocks, which employ skip connections or residual connections, as shown in Figure 2.

Figure 2. ResNet34 Network Architecture [7]

Instead of directly learning the underlying mapping, the layers within a residual block learn a residual function with respect to the block's input. This residual function is then added to the original input via the skip connection. These skip connections allow gradients to flow more directly through the network during backpropagation, facilitating the training of networks with hundreds or even thousands of layers [7].

2. Related Works

Semantic segmentation partitions an image into several semantically meaningful coherent parts

and classifies each part into one of the predetermined classes. Existing semantic segmentation methods are unreliable for autonomous driving systems, as they ignore the differing levels of importance of distinct classes for safe driving. For instance, pedestrians, cars, and cyclists in a scene are far more critical than the sky or buildings, so segmentation must be as precise as possible. To incorporate the importance of object class information, the work present in [8] develops an "Importance-Aware Loss" (IAL), which emphasizes objects critical for autonomous driving. The IAL operates under a hierarchical structure where classes with varying importance are located at different levels, thus assigning a distinct weight to them.

Studies like [9] analyze semantic segmentation failures, which are crucial for autonomous driving systems, and detect failure cases in predicted segmentation maps by calculating the Mean Intersection over Union (mIoU). Authors developed a deep neural network to predict the segmentation map's mIoU without the ground truth, and introduced a new loss function to train on imbalanced data. Safety is an extremely critical factor in autonomous driving systems, where issues related to safety metrics for artificial neural networks [10], which are a type of artificial neuron architecture with functional similarities to biological neurons [11], have been extensively studied for semantic segmentation problems. The work presented in [12] categorizes road detection algorithms into three types considering different types of information: feature-based techniques, model-based techniques, and region-based techniques. Feature-based solutions are more effective, but they require roads with well-defined and easily identifiable markings; noise can disrupt the entire detection process. Model-based techniques are more robust; however, they are severely restricted by the geometry of the models. Region-based techniques use machine learning approaches, enabling them to handle noise problems and constant changes in the environment. Research has largely overlooked the problems posed by paved roads and roads with many potholes, as well as unexpected pavement interruptions, conditions commonly found in developing countries and even in certain parts of urban areas, such as city outskirts. For an autonomous vehicle to operate successfully, it must be able to handle these conditions during its operation. The work [13] presents road marking methods (lane detection), a system that is now integrated into Advanced Driver Assistance Systems (ADAS). This work also discusses how road models are represented, presents methods for feature extraction (e.g., image intensities, edge magnitudes and orientations, comparative models), and discusses whether any type of post-processing (e.g., Hough Transform) is employed. Although published in 2013, the work [14] reviews articles published only up to 2009, considering active vision sensors (LiDAR). The review also addresses the main challenges in road detection, such as weather and lighting conditions, the presence of shadows, other

vehicles, people, or objects, and different road geometries. The work by [15] presents a review of articles published between 2005 and 2010 and also considers active vision sensors. It is a study focused on road detection, particularly lanes. Finally, authors in [16] presents work focused on road markings, concentrating on feature-based approaches like lane detection, stereoscopic analysis and edge-based segmentation.

3. Datasets for Autonomous Vehicle Navigation

Several benchmark datasets for road detection were analyzed: KITTI dataset by [17], CaRINA dataset by [18], CamVid dataset by [19], CityScapes dataset by [20], RTK dataset, and OffRoadScene dataset by [21]. One of the most widely used and cited datasets in path detection articles is the KITTI dataset, from the Karlsruhe Institute of Technology, Germany, featuring rural, urban, and highway scenarios. It presents situations with varying illumination. Most captures are on paved roads and it contains scenarios with many vehicles and pedestrians, as well as scenarios with little movement. For this work, the RTK dataset was used due to its well-curated road samples from Brazil, featuring different surface types such as asphalt variations, other types of pavements, and unpaved roads. It also includes situations with road damage (e.g., potholes).

4. Materials and methods

4.1. Dataset

The proposed method in this study addresses the semantic segmentation of road images using a neural network architecture that combines U-Net and ResNet. The objective is to classify each image pixel into one of the following categories: background, asphalt road, paved road, and unpaved road. The RTK dataset was used, which contains road images with different surface types and conditions. This dataset was chosen because it already has segmentation masks defined for each class, considerably reducing the time required for creating a database of images and segmentation masks. The dataset was divided into three parts: training, validation, and test, to evaluate the model's performance. This division was performed randomly, ensuring that each part represented the diversity of the original dataset.

4.2 Architecture

The U-Net network was combined with a ResNet backbone, which acts as an encoder to extract relevant features from the images. ResNet, proposed by researchers from Microsoft Research, uses residual blocks and skip connections to mitigate the vanishing/exploding gradient problem, enabling the training of deeper and more effective networks [22]. Different versions of ResNet (ResNet18, ResNet34, ResNet50) were explored as the model's

backbone, allowing an evaluation of the impact of network depth on road segmentation.

4.3 Preprocessing and Data Augmentation

The images were resized to a standard resolution of 512 x 512 pixels, along with data augmentation techniques such as vertical and horizontal flips (at 90º), rotations (at 30º), and {zooms (at 2x). These transformations aim to increase the variability of the training data, making the model more robust to different lighting, orientation, and scale conditions.

4.4 Model Training

Model training was initially performed with ResNet18, using the three-cycle training strategy proposed by fast.ai. This technique involves training the model on progressive image resolutions (1:8, 1:4, 1:2, and 1:1) over three phases. This procedure aims to initially train general features and shapes, followed by object features, and finally, object textures. Learning was optimized using the fast.ai's fit one cycle method, which adaptively adjusts the learning rate and momentum. The learning rate was determined using the fast.ai learning rate find function, which helps identify the optimal learning rate for training. Finally, the loss function used was Cross Entropy Loss, suitable for multi-class classification problems, along with the Adam optimizer, known for its efficiency in training deep neural networks.

4.5 Evaluation Metrics

The model's performance was evaluated using the Dice and Jaccard metrics, which measure the similarity between the predicted segmentation masks and the ground truth masks. The Jaccard Coefficient, also known as Intersection over Union (IoU), is calculated as the ratio between the area of overlap and the area of union of the predicted and ground truth masks. The Dice metric is similar to IoU but gives more weight to true positives. In semantic segmentation tasks, the objective is to classify each pixel in an image, assigning it a label that represents the class of the object to which it belongs. To evaluate how well the model performs this task, metrics are used that compare the model's predicted segmentation with the actual segmentation (or ground truth).
In the context of image segmentation, these sets are the predicted segmentation mask (the model's output) and the ground truth segmentation mask. Mathematically, the IoU is defined as:

$$IoU = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

Where:

- A represents the set of pixels belonging to a specific class in the predicted mask;
- B represents the set of pixels belonging to the same class in the ground truth mask;
- $|A \cap B|$ represents the number of pixels present in both A and B (the intersection);
- $|A \cup B|$ represents the number of pixels present in A or in B or in both (the union).

The IoU measures the ratio between the area of overlap of the predicted and ground truth segmentations and the total area covered by both segmentations. An IoU of 1 indicates perfect overlap, while an IoU of 0 indicates no overlap.

The Dice metric, also known as the Dice Coefficient, is another commonly used metric to evaluate the similarity between two images. It is very similar to IoU, but differs slightly in its formulation:

$$Dice = \frac{2 \cdot |A \cap B|}{|A| + |B|} \qquad (2)$$

Where A, B, $|A \cap B|$, $|A|$ and $|B|$ have the same meaning as in the IoU formula. The main difference between the Dice metric and IoU is that the Dice metric gives more weight to true positives (the correctly classified pixels. This is because the intersection is weighted twice as much, making it tend to be more sensitive to small details, and may be more appropriate when evaluating the segmentation accuracy of small objects.

5. Results and Conclusion

By comparing U-Net and ResNet architectures, as shown in Table 1, we demonstrated that the combination of a ResNet34 encoder with a U-Net decoder, trained on images at half the original resolution, yielded the best performance, achieving the highest Dice and IoU scores. This result indicates a robust capability for classifying pixels into essential categories such as asphalt, paved, and unpaved roads, as shown in Figures 3, 4 and 5.

Figure 3. Results for paved road



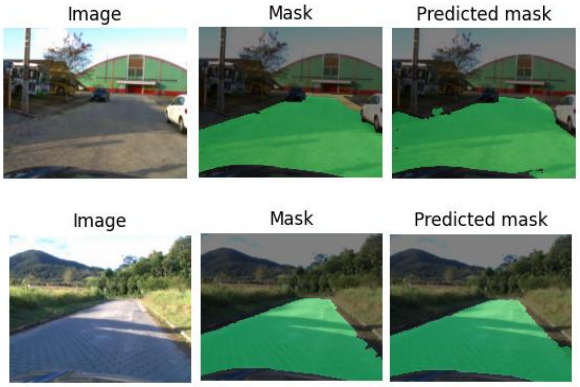


Figure 4. Results for paved and unpaved road

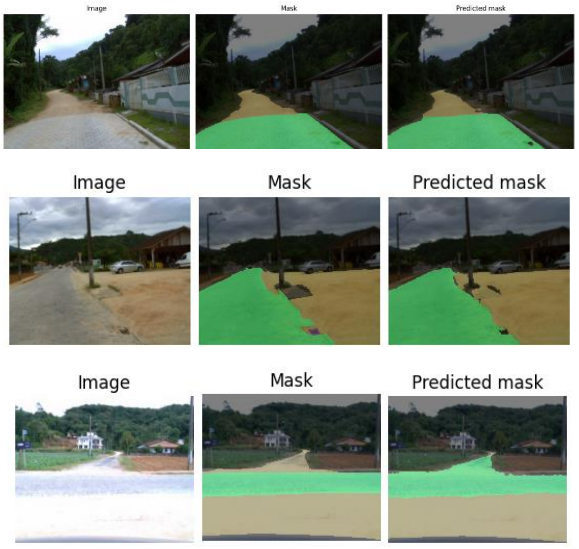


Figure 5. Results for asphalt and unpaved road

The finding that a lower image resolution produced superior results is a significant outcome of this work. By decreasing the level of detail, the model is forced to learn more general and robust features of the road, minor texture variations, or irrelevant details in the vegetation. This not only improves the model's generalization capacity for varied and poorly maintained roads but also offers a considerable practical advantage: it reduces the computational cost of both training and inference, a critical factor for real-time application in embedded systems with limited processing power. Future work will be directed towards two main fronts. First, the cross-validation of the best-performing model (ResNet34 at 1/2 resolution) on a wider range of datasets to ensure its robustness and reliability under different conditions. Second, the optimization of the architecture for deployment on embedded hardware, seeking a balance between accuracy and inference speed. As discussed in the literature, for safety-critical systems, understanding when the model is uncertain or likely to fail is as important as its accuracy. Implementing techniques to predict segmentation failures could significantly increase the safety and reliability of the autonomous navigation system, bringing it closer to practical and widespread use in challenging environments like those found in Brazil.

Table 1. Resolutions and performances

Architecture	Resolution	Dice	IoU
ResNet18	1/8	0.615	0.607
ResNet18	1/4	0.641	0.615
ResNet18	1/2	0.654	0.633
ResNet18	1/1	0.655	0.634
ResNet34	1/8	0.616	0.610
ResNet34	1/4	0.635	0.610
ResNet34	1/2	0.656	0.635
ResNet34	1/1	0.636	0.615
ResNet50	1/8	0.622	0.619
ResNet50	1/4	0.643	0.618
ResNet50	1/2	0.646	0.625
ResNet50	1/1	0.650	0.630

References

[1] CNT, Confederação Nacional do Transporte. "*Malha rodoviária total"*, 2025.

[2] D. Agyei Kyem, I. Denteh, J. Asamoah, A. A. Tutu, and C. Aboah, "*Advancing Pavement Distress Detection in Developing Countries: A Novel Deep Learning Approach with Locally-Collected Datasets*", 2024. (arXiv: 2408.05649)

[3] T. Rateke, K. Justen, V. Chiarella, A. Sobieranski, E. Comunello, A. Wangenheim. "*Passive vision region-based road detection: A literature review*". ACM Computing Surveys, 2020. 52(31):1–34.

[4] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Gläser, F. Timm, W. Wiesbeck, and K. Dietmayer. "*Deep Multi-Modal Object Detection and Semantic Segmentation for Autonomous Driving: Datasets, Methods, and Challenges*". IEEE Transactions on Intelligent Transportation Systems, vol. 22, no. 3, pp. 1341–1360, Mar. 2021, doi: 10.1109/TITS.2020.2972974.

[5] 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, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., Cham: Springer International Publishing, 2015, pp. 234–241. doi: 10.1007/978-3-319-24574-4-28. (arXiv:1505.04597 [cs.CV])

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "*ImageNet classification with deep convolutional neural networks*" in Advances in Neural Information Processing Systems 25 (NIPS 2012), F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., Curran Associates, Inc., 2012, pp. 1097–1105.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "*Deep Residual Learning for Image Recognition*" in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90. (arXiv:1512.03385)

[8] B. Chen, C. Gong, and J. Yang, "*Importance-aware semantic segmentation for autonomous vehicles*". IEEE Transactions on Intelligent Transportation Systems, vol. 20, no. 1, pp. 137–147, 2019.

[9] J. Song, W. Ahn, S. Park, and M. Lim, "*Failure detection for semantic segmentation on road scenes using deep learning*". Applied Sciences, vol. 11, no. 4, p. 1870, 2021.

[10] C. Cheng, A. Knoll, and H. Liao, "*Safety metrics for semantic segmentation in autonomous driving*" in 2021 IEEE International Conference on Artificial Intelligence Testing (AITest), 2021, pp. 57–64.

[11] F. Borba. "*Redes neurais profundas e ensemble de classificadores: uma aplicação em imagens*". Dissertação (mestrado) – Universidade Estadual de Campinas, 2021, Instituto de Matemática, Estatística e Computação Científica.

[12] Z. Shengyan, G. Jianwei, C. Huiyan, and K. Iagnemma, "*Road detection using support vector machine based on online learning and evaluation*" in IEEE Intelligent Vehicles Symposium., 2010, pp. 256–261.

[13] J. McCall and M. Trivedi, "*Video-based lane estimation and tracking for driver assistance: Survey, system, and evaluation*". IEEE Transactions on Intelligent Transportation Systems., pp. 20–37, 2006.

[14] S. Yenikaya, G. Yenikaya, and E. D ¨uven, "*Keeping the vehicle on the road: A survey on on-road lane detection systems*". ACM Comput. Surv., pp. 0–43, 2013.

[15] A. Hillel, R. Lerner, D. Levi, and G. Raz, "*Recent progress in road and lane detection: A survey*". Machine Vision and Applications., pp. 727–745, 2014.

[16] B. Shin, Z. Xu, and R. Klette, "*Visual lane analysis and higher-order tasks: A concise review*". Machine Vision and Applications., pp. 1519–1547, 2014.

[17] A. Geiger, P. Lenz, and R. Urtasun, "*Vision meets robotics: The kitti dataset*". Int. J. Rob. Res., pp. 1231–1237, 2013.

[18] P. Y. Shinzato, V. Grassi, F. S. Osorio, and D. F. Wolf, "*Fast visual road recognition and horizon detection using multiple artificial neural networks*" in IEEE Intelligent Vehicles Symposium., 2012, pp. 1090–1095.

[19] G. J. Brostow, F. Julien, and R. Cipolla, "*Road area detection based on texture orientations estimation and vanishing point detection*". The SICE Annual Conference, 2013, pp. 1138–1143.

[20] M. Cordis, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "*The cityscapes dataset for semantic urban scene understanding*" in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3213–3223.

[21] E. Shang, J. Zhao, J. Li, X. An, and T. Wu, "*Offroadscene: An open database for unstructured road detection algorithms*" in International Conference on Computer Sciences and Applications., 2013, pp. 779–783.

[22] H. He, X. Zhang, S. Ren, and J. Sun, "*Deep Residual Learning for Image Recognition*". arXiv 1512.03385, 2015.

[23] NVIDIA, "*Deep Learning Fundamentals Course*" Accessed: May 01, 2025. [Online]. Available: https://www.nvidia.com/pt-br/training/instructor-led-workshops/deep-learning-fundamentals/