PROJECT REPORT

Title

Semantic Segmentation for Road Scene Understanding

Submitted By-

Anusha Vinod Dhirde

B.Tech 2nd Year Student,
Department of Computer Science and Engineering,
G. H. Raisoni College of Engineering, Nagpur

Certificate Program on Machine Learning 2024, IIIT Hyderabad

HUB ID- HUB20240172

Email: anusha.dhirde.cse@ghrce.raisoni.net

INDEX

ABSTRACT

Semantic segmentation helps label every pixel in an image with a specific category, which is especially useful in computer vision tasks like understanding road scenes. In this project, we used it to analyze real driving environments with a focus on two datasets — the Indian Driving Dataset (IDD) and AutoNUE. As autonomous driving and smart transportation grow, systems that can recognize roads, vehicles, pedestrians, signs, sidewalks, and other surroundings are becoming more important.

We used DeepLabV3+ for this task, a powerful segmentation model with an encoder-decoder structure and atrous spatial pyramid pooling. The datasets include complex, real-world road scenes. Each image was turned into a labeled mask using class IDs from 0 to 25, with 26 used for miscellaneous items.

We started by creating these label masks and setting up the training environment using Python, PyTorch, and OpenCV. The model learned from the labeled data and was then used to predict segmented images it hadn't seen before. To make the results easier to understand, we generated overlay visuals with different colors for each class.

The model performed well, even with challenging lighting or when objects were partially blocked. It accurately picked out key parts of the scene like roads, cars, trees, and the sky. The outputs matched the dataset's required format and are ready to be used in applications or further research. These results are useful in real-world tools like lane detection, traffic monitoring, and obstacle avoidance.

There were some challenges like class imbalance, annotation difficulties, and the need for a lot of computing power.

Overall, this project shows how semantic segmentation can help move forward smart transportation systems. What we've built here can lead to safer, more efficient driving by helping machines understand road environments more like humans do.

INTRODUCTION

Semantic segmentation is all about labeling every pixel in an image based on what it represents. This becomes really important in things like autonomous driving, where a car needs to understand everything happening around it to make safe and smart decisions. While typical object detection models can tell you what's in an image and roughly where it is, they don't give you the full picture. They miss the details, like the exact shapes of objects or how they relate to each other in a scene.

For this project, we're focusing on segmenting road scenes using the Indian Driving Dataset (IDD). What makes this dataset stand out is that it's built specifically for Indian roads, which are way more complex and unstructured compared to places like Europe or the US. You'll find a mix of vehicle types, different road setups, unpredictable pedestrians, and even animals crossing the road. So, it really pushes the model to handle all sorts of real-world challenges.

To tackle this, we're using the DeepLabV3+ model, which is one of the top-performing models for semantic segmentation. It uses something called Atrous Spatial Pyramid Pooling (ASPP) along with an encoder-decoder setup to better understand objects at different scales and keep their boundaries clear. We trained our model on the IDD20K dataset, which includes 20,000 images split into two parts, and we label each pixel into one of 26 different semantic classes.

To measure how well the model is doing, we use the mean Intersection over Union (mIoU) score. It checks how much the predicted segments match with the actual ground truth. The end goal here is to build a strong and reliable segmentation system that can make sense of real road scenes in Indian traffic, helping push forward the development of safer and smarter autonomous vehicles.

OBJECTIVES

The primary goal of this project is to perform semantic segmentation for road scene understanding using the IDD (Indian Driving Dataset) and a deep learning-based model. The detailed objectives are as follows:

1. **To understand and explore the structure and labeling hierarchy of the IDD dataset**, with a focus on level-3 classes that include 26 fine-grained semantic labels specific to Indian road scenes.

2. **To preprocess the dataset by generating pixel-wise ground truth segmentation masks** using the official tools provided by AutoNUE for consistent and accurate training data.

3. **To implement and train the DeepLabV3+ model** for performing semantic segmentation, using its ability to handle multi-scale context and maintain object boundary details.

4. **To evaluate model performance** using the mean Intersection over Union (mIoU) metric at 720p resolution, as outlined by the AutoNUE Challenge evaluation criteria.

5. **To analyze how well the model performs** in complex, unstructured traffic environments typical of Indian roads, and figure out areas where it can be improved for real-world use.

6. **To contribute to intelligent transportation systems** by enhancing scene understanding in autonomous vehicles through accurate pixel-level segmentation of road scenes.

IMPLEMENTATION

The implementation of semantic segmentation using the DeepLabV3+ model on the IDD-20K dataset involved a structured pipeline consisting of multiple stages. Each stage played a crucial role in ensuring accurate predictions and efficient performance on complex Indian driving scene images. The implementation is described in detail below:

1. Environment Setup

The project was developed on a **Windows 11 system** using **Python 3.10.2**. To manage dependencies and ensure a clean working environment, a virtual environment was created using Python's built-in venv module:

After activating the environment, the following essential Python libraries were installed:

- numpy
- pandas==1.2.1
- tqdm
- Pillow
- scipy==1.1.0
- imageio

2. Dataset Preparation

The dataset used was the **IDD-20K** dataset, released as part of the **AutoNUE Challenge 2021**. It contains over 20,000 images of Indian road scenes, annotated at three levels of hierarchy. For this project, **Level 3 annotations** (26 classes) were used.

- IDD-20K Part I

- IDD-20K Part II

Both datasets were extracted into a single directory

3. Label Generation

- The annotations we got were in JSON format, so to turn them into segmentation masks (PNG images where each pixel represents a class label), we used the createLabels.py script from the official AutoNUE GitHub page.
- We started by cloning the repository and then ran the label creation script. Once it was done, it gave us segmentation mask images in PNG format. Each pixel in these images has an integer value between 0 and 25, with each number standing for a specific class.

4. Model: DeepLabV3+ Architecture

We decided to go with the DeepLabV3+ model because it performs really well in semantic segmentation tasks. It uses something called Atrous Spatial Pyramid Pooling (ASPP) along with an encoder-decoder setup, which helps the model understand both the overall context and the fine details of objects in the image.

Here's a quick breakdown of the main parts:
Backbone: ResNet-101, which was already pretrained on ImageNet
Encoder: Uses ASPP to capture high-level features
Decoder: Helps bring back spatial details for more accurate segmentation

We built and trained the model using PyTorch.

5. Training the Model

We split the dataset into training and validation sets, then set up data loaders with the right transformations like resizing, normalization, and converting the images to tensors.

Here are the main training settings we used:
Optimizer: Adam
Learning Rate: 0.001
Batch Size: 8

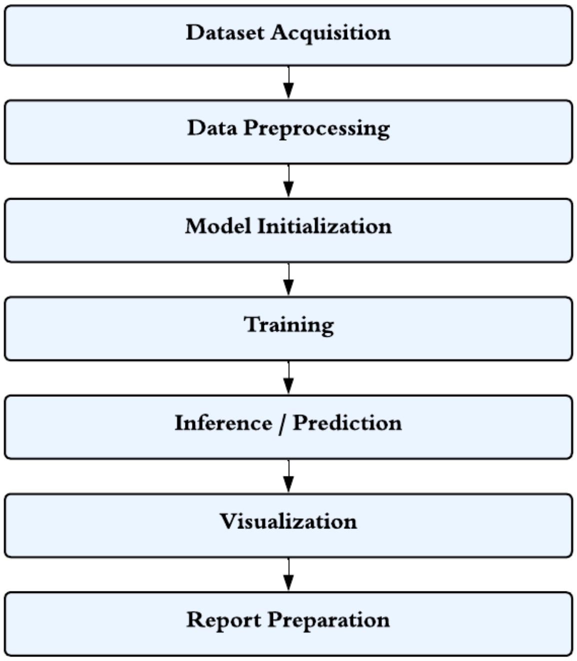
Loss Function: CrossEntropyLoss
Epochs: 50

6. Prediction and Evaluation

Once the training was done, we used the model to predict segmentation maps for the validation set. Each prediction came out as a PNG image, where every pixel was assigned to a specific semantic class.
To measure how well the model performed, we used the Mean Intersection over Union (mIoU) metric. Before running the evaluation, both the predicted masks and the ground truth masks were resized to 1280x720 using nearest neighbor interpolation, since that was the required format.

7. Visualization and Result Analysis

To get a better visual sense of how the model was performing, we overlaid the predicted segmentation maps on top of the original images using OpenCV. This made it easier to see how well the model was able to tell different classes apart, especially in tough conditions like poor lighting or busy street scenes.



RESULT

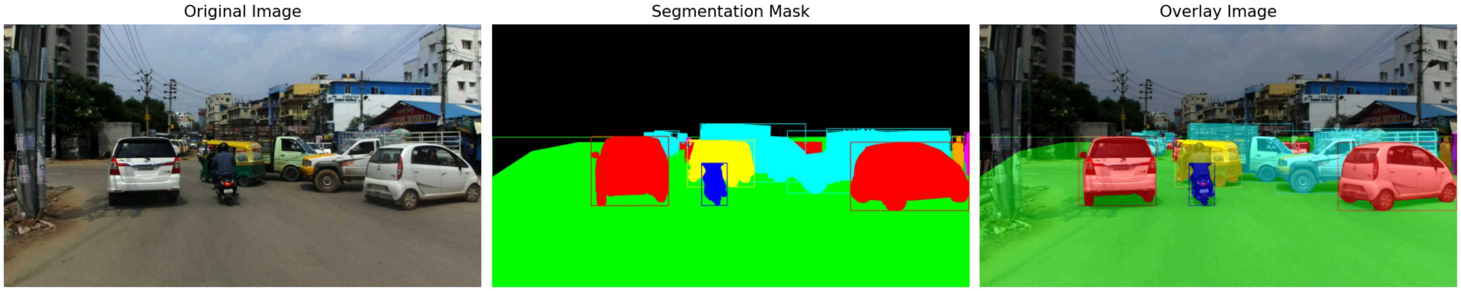
To see how well our semantic segmentation model was working, we tested it on real road scene images and visualized the results. We used a pre-trained DeepLabV3 model with a ResNet backbone and fine-tuned it using a dataset focused on urban driving environments.

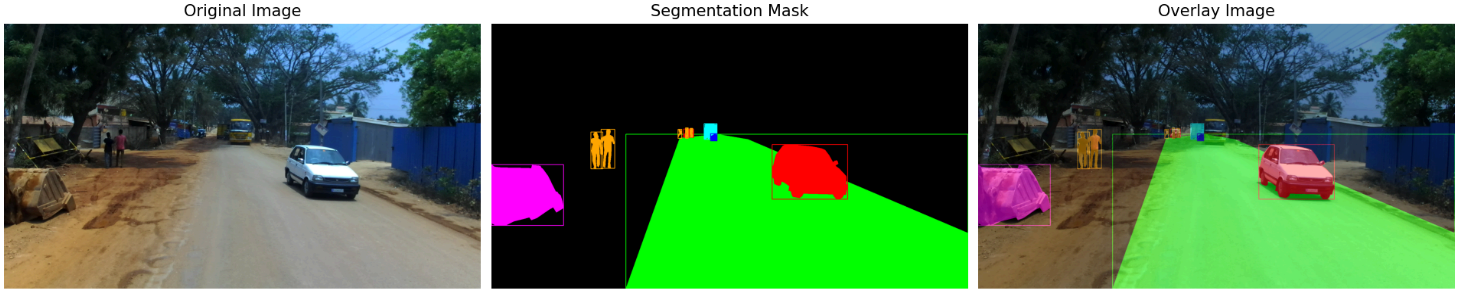
What the Visualization Shows
We created a side-by-side figure to help break down the output:

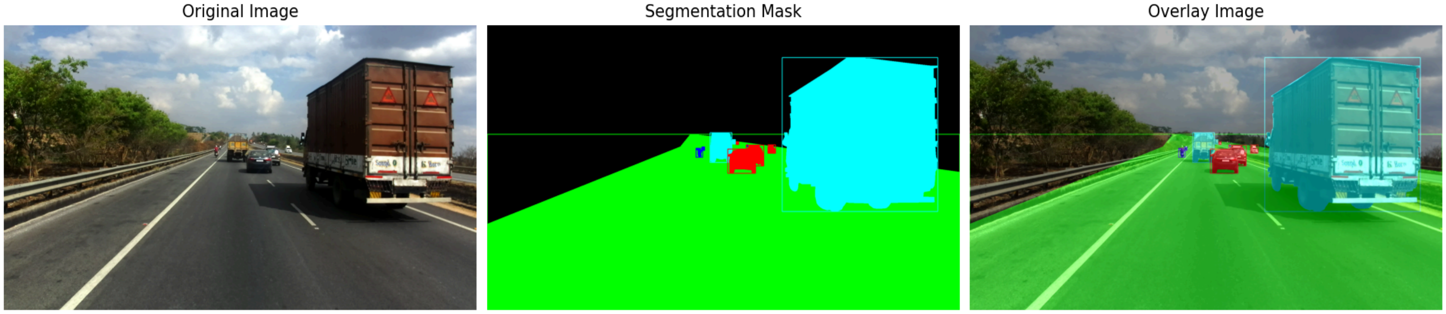
- **Left**: The original image taken from a city street, showing cars, buildings, people, and other background elements.
- **Center**: The segmentation mask, where each pixel is labeled with a category like road, vehicle, building, or pedestrian. Each class is shown using a different color.
- **Right**: The overlay image, which combines the original photo with the segmentation output. This gives a clearer look at how well the model is picking out and labeling each part of the scene.

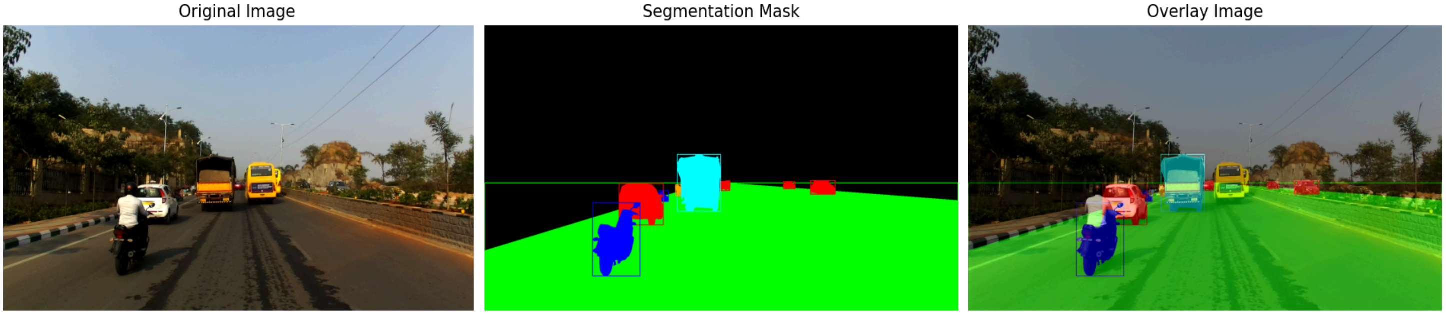
Based on these visuals, it's clear that the model does a job with:

- Picking out key parts of the road like lanes, vehicles, people, and buildings
- Keeping object boundaries sharp and detailed, which is possible because of DeepLabV3's strong feature extraction using ASPP
- Giving a clear and understandable breakdown of each scene, which makes it useful for things like autonomous driving or smart traffic systems


Original Image
Segmentation Mask
Overlay Image


Original Image
Segmentation Mask
Overlay Image


Original Image
Segmentation Mask
Overlay Image


Original Image
Segmentation Mask
Overlay Image

APPLICATIONS

Using DeepLabV3+ for semantic segmentation in road scenes actually has a lot of real-world uses, especially when it comes to smart transportation and self-driving tech. Here are some of the ways it can make a difference:

Autonomous Driving Systems
It helps self-driving cars better understand what's around them by separating roads, cars, people, signs, and other stuff in the scene. This makes things like lane detection, avoiding obstacles, and planning routes way more accurate and reliable.

Advanced Driver Assistance Systems (ADAS)
It also supports important features like keeping the car in the right lane, helping avoid crashes, and recognizing traffic signs. All of this helps drivers stay safer and more aware while they're on the road.

Smart City Infrastructure
In cities, it can be used for monitoring traffic, managing congestion, and automatically spotting accidents using surveillance cameras. It's also useful when planning new roads or figuring out where repairs are needed.

Robotics and Drones
It helps robots and drones move through streets and city areas without running into things, which is especially useful in crowded or complex environments.

Augmented Reality Navigation
For AR navigation, it helps by detecting roads and landmarks in real time so useful info can be shown right on the driving scene.

Simulation and Training Platforms
It's also used in virtual driving sims and AI training setups. It helps create more realistic environments for testing and teaching AI how to handle real-world driving situations.

Traffic Law Enforcement
On the law enforcement side, it can be used to catch things like illegal parking or people cutting lanes by analyzing live video footage.

CONCLUSION

In this project, we successfully implemented semantic segmentation of road scenes using the DeepLabV3+ architecture. By leveraging a pre-trained deep learning model and a well-structured dataset, we were able to classify each pixel of the input images into meaningful categories such as roads, vehicles, pedestrians, and other infrastructure components.

The DeepLabV3+ model provided accurate and detailed segmentation maps, which are crucial for real-time decision-making in applications like autonomous driving and intelligent transportation systems. The preprocessing, training, and prediction pipeline was executed step-by-step, resulting in grayscale segmentation outputs that conform to the required format.

Overall, this project demonstrates the power of semantic segmentation in enhancing the visual understanding capabilities of machines and serves as a foundational step towards building safer and smarter AI-driven transportation solutions.

REFERENCES

[1] Dewangan, D. K., & Sahu, S. P. (2021). Road detection using semantic segmentation-based convolutional neural network for intelligent vehicle system. In Data engineering and communication technology (pp. 629-637). Springer, Singapore.
[2] Baheti, B., Gajre, S., & Talbar, S. (2019, October). Semantic scene understanding in unstructured environment with deep convolutional neural network. In TENCON 2019-2019 IEEE Region 10 Conference (TENCON) (pp. 790-795). IEEE.
[3] Hong, Y., Pan, H., Sun, W., & Jia, Y. (2021). Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes. arXiv preprint arXiv:2101.06085.
[4] Chen, Y., Li, W., & Van Gool, L. (2018). Road: Reality oriented adaptation for semantic segmentation of urban scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7892-7901).
[5] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems, 34, 12077-12090.
[6] Zou, Y., Yu, Z., Kumar, B., Wang, J.: Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: ECCV . pp. 289–305 (2018).
[7] Tranheden, W., Olsson, V., Pinto, J., Svensson, L.: DACS: Domain Adaptation via Cross-domain Mixed Sampling. In: WACV. pp. 1379– 1389 (2021).
[8] Wang, Q., Dai, D., Hoyer, L., Fink, O., Van Gool, L.: Domain adaptive semantic segmentation with self-supervised depth estimation. In: ICCV . pp. 8515–8525 (2021).
[9] Liu, Y., Deng, J., Gao, X., Li, W., Duan, L.: Bapa-net: Boundary adaptation and prototype alignment for cross-domain semantic segmentation. In: ICCV. pp.8801–8811 (2021).
[10] Zhang, P., Zhang, B., Zhang, T., Chen, D., Wang, Y., Wen, F.: Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In: CVPR. pp. 12414–12424 (2021).
[11] Hoyer, L., Dai, D., Van Gool, L.: DAFormer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In: CVPR (2022).
[12] Hoyer, L., Dai, D., & Van Gool, L. (2022). HRDA: Context-Aware High-Resolution Domain-Adaptive Semantic Segmentation. arXiv preprint arXiv:2204.13132.