EPFL CS-433-Machine Learning Road Segmentation Project Report

Efe Tarhan
School of Computer and Communication Sciences, EPFL, Switzerland

***Abstract*—This report presents a comprehensive study on road segmentation using deep learning models. The task focuses on identifying road surfaces within high-resolution aerial imagery, a critical component in applications such as autonomous driving and urban planning. Traditional models like U-Net and PSPNet were evaluated and compared against the Transformer-based SegFormer architecture. SegFormer demonstrated superior performance, achieving an F1 score of 0.891 due to its hierarchical feature extraction and adaptability to environmental occlusions. Data augmentation techniques, including black box augmentation, significantly improved generalization and robustness, increasing F1 scores across all models by approximately 0.1. The results highlight the potential of combining advanced architectures with effective preprocessing strategies for robust road segmentation.**

I. Introduction

Road segmentation is a critical task in computer vision, focusing on identifying road surfaces within images. It underpins applications such as autonomous driving and urban planning. In autonomous driving, accurate segmentation allows vehicles to navigate safely by identifying drivable areas, enhancing road safety and reducing traffic incidents [1]. In urban planning, it supports mapping infrastructure, optimizing traffic flow, and promoting sustainable development through efficient resource allocation [2].

Deep learning models like U-Net [1], PSPNet [2], and Transformer-based models such as SegFormer [3] have advanced road segmentation by leveraging hierarchical feature extraction and contextual understanding. Recent innovations, including generative pre-trained models like SegGPT [4], address challenges like varying road conditions, occlusions, and diverse terrains. However, robust segmentation in high-resolution aerial images remains complex due to the heterogeneous and intricate nature of the data.

High-resolution aerial images pose unique challenges, including variations in road width, occlusions, lighting changes, and spatial resolution differences, which complicate model performance and generalization. Traditional convolutional models often struggle with these issues, whereas emerging Transformer-based models offer improved robustness and adaptability. This work addresses these challenges using SegFormer, a hierarchical Transformer-based model, to achieve accurate and efficient road segmentation under diverse conditions.

The primary objectives of this work are as follows:

- To develop a robust road segmentation model capable of processing high-resolution aerial imagery.
- To compare the performance of the used model with classical models such as U-Net and PSPNet.
- To provide qualitative and quantitative analyses of the segmentation performance, highlighting the strengths and limitations of the selected approach.

II. Methodology

A. Dataset Description

1) Training Set: The training set contains 100 aerial images of size 400×400 pixels in RGB format. Each image captures a section of terrain from an aerial perspective, including roads, vegetation, buildings, and other urban or natural features.

For each training image, there is an accompanying binary mask of the same size (400×400 pixels). These masks are ground truth annotations that indicate the location of roads within the image. In the masks, pixels labeled as 1 represent roads, while pixels labeled as 0 correspond to non-road areas. The pair of training images and their corresponding masks provide the supervised learning data for model training. An example data from the set can be seen in Figure 1.

Figure 1. Example of a training image (left) and its corresponding binary mask (right). Roads are labeled as 1 (white), while non-road areas are labeled as 0 (black).

2) Test Set: The test set consists of 50 aerial images, each of size 608×608 pixels in RGB format. These images cover a different set of terrains and are used to evaluate the generalization performance of the trained model. Unlike the training images, the test images are larger, providing a robust test for the model's ability to adapt to input data of varying resolutions.

3) Preprocessing and Data Augmentation: To enhance the model's robustness and ensure consistency in input data, a combination of data augmentation and preprocessing techniques was applied to the dataset.

Data Augmentations: The training data was augmented using a set of transformations implemented with the `Albumentations` library. These transformations introduce variability in the training data to prevent overfitting and improve the model's generalization capabilities. The specific transformations applied are as follows: **Horizontal and Vertical Flips** were applied with a 50% probability to each training image, adding diversity to the data. **Rotation** was performed by randomly rotating images within a range of $[-90°, 90°]$, simulating different orientations of aerial views. **Brightness and Contrast Adjustments** were applied to account for variations in lighting conditions. **Gamma Transformation** was also applied enhancing the dataset's dynamic range. **Color Jitter** introduced random changes in brightness, contrast, saturation, and hue to simulate varying color conditions. Finally, **Coarse Dropout** was used to randomly mask small rectangular regions of the image, simulating occlusions and forcing the model to focus on the surrounding context.

The **SegformerImageProcessor** from the Hugging Face library was used to preprocess the data for the SegFormer model [5]. This ensured input images were normalized and resized to the required dimensions.

B. SegFormer Architecture

SegFormer is a state-of-the-art semantic segmentation model combining a hierarchical Transformer-based encoder with a lightweight MLP decoder, balancing high performance and computational efficiency for diverse segmentation tasks [3].

1) Encoder: Mix Transformer (MiT): The encoder, MiT, is a hierarchical structure that uses self-attention mechanisms to capture multiscale features, enabling SegFormer to model both global and local dependencies effectively. Unlike traditional Transformers, MiT omits positional encodings, enhancing its adaptability to varying input sizes. The encoder processes input images in stages, progressively downsampling and extracting deeper features, making it well-suited for segmenting objects at various scales in aerial imagery [3].

2) Decoder: Lightweight MLP Design: SegFormer's decoder is a simple MLP that aggregates multilevel features from the encoder to produce the segmentation map. Despite its minimalistic design, the decoder effectively combines global context with local details, leveraging the rich features extracted by the encoder. This efficiency sets it apart from more complex architectures like U-Net or PSPNet [3].

3) Key Advantages: SegFormer introduces several key innovations:

- Positional Encoding-Free Design: Omits positional encodings to ensure robust performance across varying image resolutions.
- Multiscale Feature Extraction: The hierarchical encoder captures features at multiple scales, crucial for tasks like aerial road segmentation where objects vary in size.
- Efficiency: A lightweight MLP decoder reduces computational demands while maintaining high accuracy, suitable for real-time applications [3].

Figure 2 illustrates SegFormer's architecture, highlighting the flow from multiscale feature extraction in the encoder to efficient decoding for final segmentation.

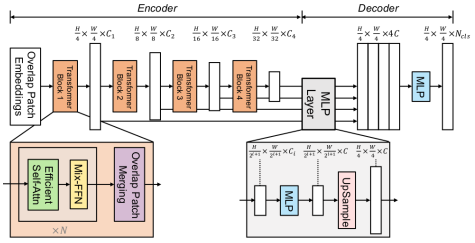

Figure 2. Overview of the SegFormer architecture, showing the Mix Transformer (MiT) encoder and the lightweight MLP decoder [3].

C. Training Setup

The SegFormer model used in this work is based on the pre-trained "nvidia/segformer-b4-finetuned-ade-512-512" [5] checkpoint, obtained from the Hugging Face Model Hub. The model has been fine-tuned for the binary road segmentation task, where the labels are defined as "road" and "no_road". The implementation was carried out using PyTorch on an NVIDIA GeForce RTX 3090 GPU with 25 GB of memory, leveraging CUDA for acceleration. A random train-validation split was applied to the dataset, allocating 90% of the data for training and 10% for validation. Due to computational limitations, k-fold cross-validation was not feasible.

1) Hyperparameters: The training setup employs the following hyperparameters: The **Number of Epochs** is set to 150. The **Learning Rate** is initialized at 1×10^{-3} and is adjusted dynamically during training using a linear scheduler. A **Learning Rate Scheduler** is used with a step size of 15 epochs and a decay factor (γ) of 0.5.

2) Optimizer: An AdamW optimizer [6] is used to update model parameters during training. This optimizer is well-suited for Transformer-based models, as it effectively handles weight decay using L2 regularization, preventing overfitting and ensuring stable convergence.

The learning rate scheduler reduces the learning rate by a factor of 0.5 every 15 epochs, improving training stability in later epochs.

3) Loss Function: Binary cross-entropy (BCE) loss is used to measure the divergence between predicted probabilities and ground truth masks, helping the model distinguish road pixels from non-road pixels.

4) Software Tools: The training pipeline utilized with GPU acceleration, leveraging the NVIDIA GeForce RTX 3090 GPU.

D. Post-Processing Methods

1) Test-Time Augmentation (TTA): TTA improves model predictions by evaluating augmented inputs during inference. We applied 90°, 180°, and 270° rotations to the input data and used majority voting across predictions to produce the final output. This approach enhances robustness to orientation variations while slightly increasing computational cost.

2) Morphological Operations: Morphological operations, including closure, opening, and dilation, were applied to refine predictions. Closure filled small gaps, opening removed noise, and dilation expanded boundaries for better connectivity. These enhancements improved precision and ensured visually coherent results.

III. Experiments

A. Traditional Models

UNet [1] and PSPNet [2] were utilized as traditional models to evaluate performance and establish a foundation for further development. Both models were trained for 100 epochs with a learning rate of 1×10^{-3}.

1) PSPNet: The following figures present the results for PSPNet. The training BCE loss curve in Figure 3, the validation F1 scores in Figure 4, and sample predictions generated by the model in Figure 5.

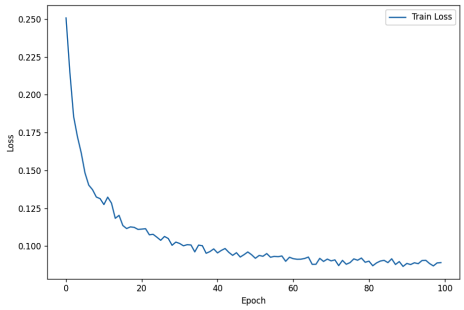

Figure 3. Training loss values for PSPNet over epochs.

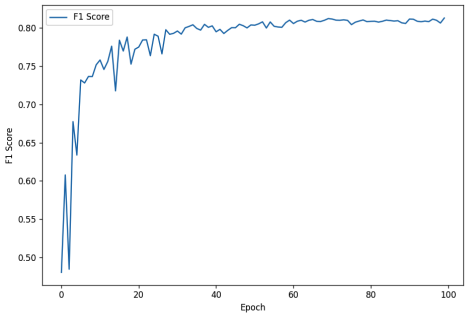

Figure 4. F1 scores for PSPNet for validation data on each epoch.

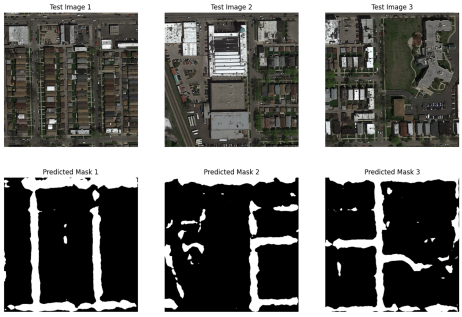

Figure 5. Sample predictions generated by PSPNet.

2) UNet: The following figures present the results for UNet with a ResNet-34 encoder, trained for 100 epochs using a learning rate of 1×10^{-4}. These include the training BCE loss curve in Figure 6, the validation F1 scores in Figure 7, and sample predictions generated by the model in Figure 8.

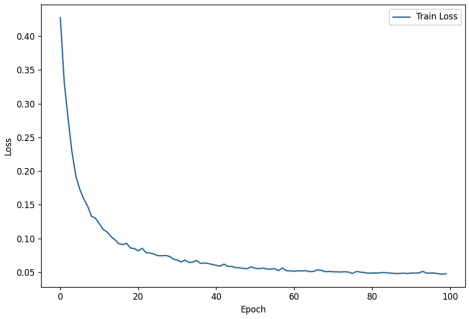

Figure 6. Training loss values for U-Net over epochs.

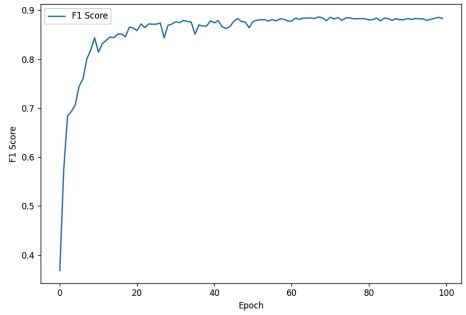

Figure 7. F1 scores for U-Net for validation data on each epoch.

From the results, it can be concluded that U-Net outperformed PSPNet both visually and in terms of metrics such as F1 scores and the reduction in training loss. Given these observations, the SegFormer model will now be employed with the expectation of achieving even better results.

B. SegFormer Performance

The SegFormer model of the Hugging Face has been trained according to the techniques described in Section II-C. The figures include the training BCE loss curve in Figure 9,

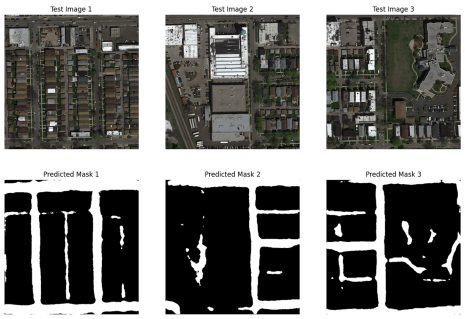

Figure 8. Sample predictions generated by U-Net.

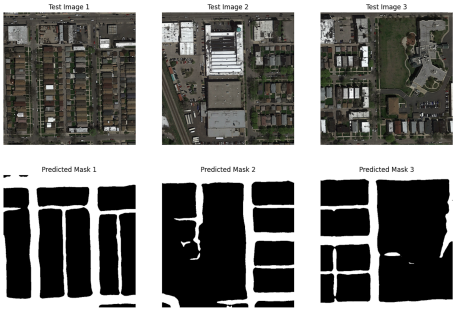

Figure 11. Sample predictions generated by SegFormer, demonstrating the model's segmentation accuracy and visual quality.

the validation F1 scores in Figure 10, and sample predictions generated by the model in Figure 11.

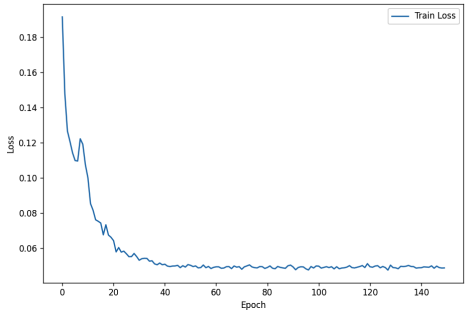

Figure 9. Training loss curve for SegFormer, showing model convergence over epochs.

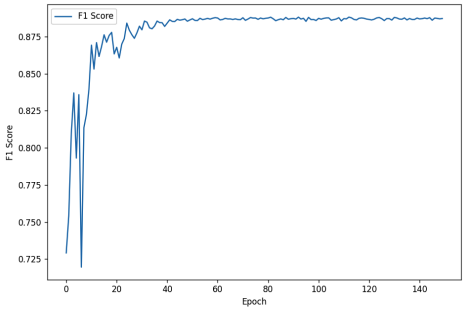

Figure 10. Validation F1 scores for SegFormer, highlighting its performance improvement over time.

From the results of PSPNet, U-Net, and SegFormer, it is visually evident that SegFormer excels in maintaining sharp and well-defined edges for roads while effectively separating buildings that share a similar color with the roads. These qualitative observations will be elaborated upon in the next section, which provides a detailed qualitative analysis of the results.

IV. Results & Analysis

The table I summarizes the performance and hyperparameter configurations for PSPNet, U-Net, and SegFormer models. Each model was trained under similar conditions, including a learning rate of 1×10^{-4}, a step size of 15, and a gamma decay factor of 0.5. However, the SegFormer model outperformed both PSPNet and U-Net, achieving the highest F1 score of 0.891. This section discusses the reasons behind these results and the impact of model architecture and data augmentation techniques on performance.

Table I
Model Comparison

Parameters	PSPnet	U-Net	Segformer
# of Epochs	100	100	150
Learning Rate	1e-4	1e-4	1e-4
Step Size	15	15	15
Gamma	0.5	0.5	0.5
F1-Score	0.813	0.883	**0.891**

While **PSPNet** achieved an F1 score of 0.813, U-Net surpassed it with an F1 score of 0.883. **U-Net's** superior performance can be attributed to its effective encoder-decoder structure, which is well-suited for capturing fine-grained details. PSPNet, although robust in extracting global context with its pyramid pooling module, struggled to handle intricate road structures effectively. **SegFormer** achieved the best F1 score of 0.891, outperforming both traditional models. Its hierarchical Transformer-based encoder enabled it to capture multiscale features and model long-range dependencies, making it particularly effective at identifying complex road structures. Moreover, SegFormer demonstrated strong adaptability to environmental occlusions, such as shadows, vegetation, and buildings, which often challenge traditional convolutional models.

The use of targeted data augmentation significantly improved the models' performance. Coarse dropout, which masks small regions of input images, played a critical role in preventing overfitting by encouraging the models to focus on contextual information rather than specific pixel patterns. This approach was particularly effective for SegFormer, leveraging its ability to capture long-range dependencies. Additionally, other augmentations such as random flips, rotations, brightness/contrast adjustments, and color jitter introduced variability into the training data, enabling the models to learn robust features that accounted for diverse environmental conditions. Collectively, these augmentations increased the F1 score by approximately 0.1 for each model.

V. Ethical Risks

In the context of road segmentation using aerial satellite images, a primary risk lies in the model's potential inability to accurately classify roads, especially in areas with visual complexity such as dense urban environments or rural regions with occlusions (e.g., shadows, vegetation, or buildings). This risk directly affects stakeholders such as urban planners, transportation authorities, and disaster response teams who depend on accurate road maps for critical planning and decision-making. Misclassification could result in flawed urban planning, resource misallocation, or delays in emergency responses, amplifying the importance of mitigating this risk.

The risk evaluation process primarily relied on the F1 score as a performance metric, balancing precision and recall to assess the model's ability to correctly identify roads. With a dataset of only 100 images, the limited size introduced challenges related to generalizability and potential bias, as the dataset might not sufficiently represent diverse road conditions, environments, or occlusion scenarios. Literature on segmentation challenges and the limitations of small datasets informed this analysis, highlighting the increased likelihood of overfitting and the reduced robustness of the model under real-world conditions.

To mitigate these risks, various strategies were employed. Data augmentation techniques were used to artificially increase dataset diversity by simulating variations in lighting, rotation, and occlusions, thereby improving the model's robustness to unseen scenarios. Additionally, the segmentation pipeline was iteratively refined based on the F1 score performance, with targeted adjustments to address observed weaknesses, such as poor performance on narrow or intersecting roads. Despite these efforts, the small dataset size posed an inherent limitation, making it difficult to comprehensively capture the variability present in real-world road networks.

The results of the risk analysis led to specific project adaptations, including a focus on maximizing the utility of the small dataset through augmentation and careful validation to ensure the F1 score improvement was not the result of overfitting. However, limitations in dataset size and computational resources remain barriers to full risk mitigation. Moving forward, expanding the dataset and incorporating additional evaluation metrics would be crucial to enhancing the reliability and applicability of the segmentation outputs for the stakeholders involved.

References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2015, pp. 234–241.

[2] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 2881–2890.

[3] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," 2021. [Online]. Available: https://arxiv.org/abs/2105.15203

[4] W. Wang, E. Xie, X. Li, X. Liu, D. Chen, Z. Wang, T. Lu, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Seggpt: Segmenting everything in context," *arXiv preprint arXiv:2304.03284*, 2023.

[5] Hugging Face, "Segformer," 2023, accessed: Dec. 14, 2024. [Online]. Available: https://huggingface.co/docs/transformers/model_doc/segformer

[6] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019. [Online]. Available: https://arxiv.org/abs/1711.05101

VI. Appendix

A. Statement of Contribution

Due to the unbalanced workload shared by the contributors of this project, this statement of contribution is considered to be written. The following are the contributions to this project by each team member:

1) ***Sabri Yiğit Arslan****:*

- He has done the necessary literature review for the project.
- He has helped the other team members through project for hyper parameter optimization and debugging problems.
- He has written the necessary sturctural files for GitHub repository and organized it.(Requirements, Readme, folder structures etc.)
- He has helped editing the report and giving it its final shape.
- He has evaluated all written functions in GitHub if they are working independent of the environment.

2) ***Efe Tarhan****:*

- He has done the literature review and tried to implement different Github repositories.
- He has written the notebooks for training the baseline models and the final segformer model.
- He has written the functions for running the main run.py() file for the submission.
- He has written all chapters of the report.
- He has obtained the best f1 score in the submission which is 0.862 and does 16 submission on the AICrowd system.
- He has helped preparing the GitHub repository with the file and folder structures.
- He has done the hyperparameter selection and optimization for the segformer.

3) ***Eren Akçanal****:*

- He has done the literature review by reading articles and finding GitHub repositories.
- He has implemented UNet individually and obtained results.
- He has found an additional dataset and tried to train one of the current models (segformer) in a notebook prepared by the team.
- He has done 5 individual submissions with his Unet and obtained f1 scores around 0.79 in the same period he has done the trainings.