Optimizing Deep Learning Architectures for High-Resolution Road Segmentation in Aerial Imagery

Maria Cherchouri, Elsa Heitz and Antoine Tissot-Favre
CS-433, EPFL Lausanne, Switzerland

***Abstract*—Accurate road segmentation in high-resolution aerial imagery supports navigation, planning, and emergency response. This work compares state-of-the-art models U-Net, SegNet, and LinkNet variants using diverse data augmentation (rotations, brightness shifts, noise) and an additional dataset to reduce biases. Among tested loss functions (e.g., BCE, Squared Dice), Binary Cross-Entropy combined with SegNet delivered the best performance, achieving an F1 score of 0.87 and an accuracy of 0.94. The findings highlight effective architectural, augmentation, and optimization strategies for robust and reliable road segmentation.**

I. Introduction

Road segmentation from high-resolution satellite and aerial imagery is crucial for applications in transportation systems, urban planning, and disaster response. Accurate road identification supports tasks like navigation, vehicle tracking, and land-use analysis but remains challenging due to environmental variability and complex urban layouts.

Deep learning has significantly advanced road segmentation with encoder-decoder architectures, such as Fully Convolutional Networks (FCN) and U-Net, which extract multi-scale features for pixel-level classification [1]. Innovations like residual blocks and attention mechanisms have further enhanced performance.

Loss functions, including binary cross-entropy and dice loss, are equally critical as they guide training and impact accuracy [2]. However, their effectiveness varies depending on datasets and tasks.

This paper evaluates state-of-the-art architectures, compares their performance, and outlines tuning strategies to adapt models for robust and accurate road segmentation.

II. Dataset Description

This study uses a dataset tailored for road segmentation, consisting of satellite and aerial imagery divided into training and testing subsets.

A. Training Dataset

The training set includes 100 RGB satellite images, each 400×400 pixels, sourced from Google Maps. Ground-truth masks accompany these images, where each pixel is labeled as: **white** for road regions and **black** for background (non-road).

These binary masks enable supervised learning for pixel-wise road segmentation (Fig. 1).

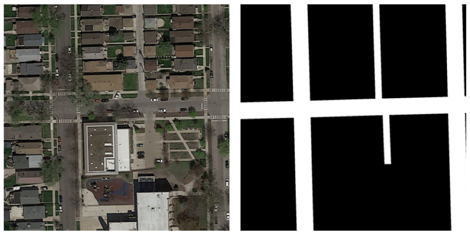

Figure 1. Image (left) and corresponding ground truth (right).

B. Testing Dataset

The testing set consists of 50 RGB images, each 608×608 pixels. Ground-truth masks are withheld during testing, requiring models to predict road regions for evaluation.

C. Dataset Summary

- **Training set:** 100 images (400×400) with binary ground-truth masks.
- **Testing set:** 50 images (608×608) without ground-truth masks.

The variation in image size between training and testing presents a challenge, requiring models to generalize well for accurate predictions on larger test images.

III. Data Augmentation and Preparation

To improve model robustness and generalization, we implemented a data preparation pipeline addressing biases in the original dataset and incorporating an external dataset to represent underrepresented features like parking lots and highway entries.

A. Incorporation of Additional Dataset

We incorporated an external dataset from Lucci et al. in *Learning Aerial Image Segmentation from Online Maps*, discovered through shared project code. The dataset features diverse urban and rural scenes from cities like Chicago, Zurich, Berlin, Paris, and Tokyo, closely aligning with our test set's characteristics [3].

B. Data Generation and Preprocessing

The dataset was generated as follows:

- **Image Source:** Aerial RGB images from Google Maps; road/building labels from OpenStreetMap (OSM).

- **Label Creation:** Roads were approximated from center lines with category-based widths, and buildings were derived from corner polygons. The labels were modified to align with our dataset's labeling scheme, retaining only two classes: road and background.
- **Preprocessing:** Images were resized and augmented using rotations, flips, and brightness adjustments (Section 2.1).

This dataset improved generalization to diverse road structures.

C. Data Augmentation Techniques

To address road orientation biases (Fig. 2), the following augmentations were applied:

- **Rotations:** Fixed rotations $(15°, 30°, 45°, 60°, 90°, 180°, 270°)$ and random rotations were applied to diversify road orientations.
- **Flipping:** Random horizontal and vertical flips simulated mirrored scenarios.
- **Padding:** To ensure consistent input dimensions after transformations, images were padded as necessary, maintaining the aspect ratio and preventing information loss.
- **Brightness and Contrast Adjustment:** Brightness and contrast were varied randomly between 0.8 and 1.2 to emulate different lighting conditions.
- **Gaussian Noise:** Added Gaussian noise (mean 0.0, std 0.08) to simulate sensor imperfections.
- **Random Transformations:** Combined augmentations further increased dataset diversity.

D. Summary

By augmenting the dataset and integrating a diverse external dataset, we mitigated biases and improved the model's ability to generalize to complex scenarios, including urban roads and highway entries.

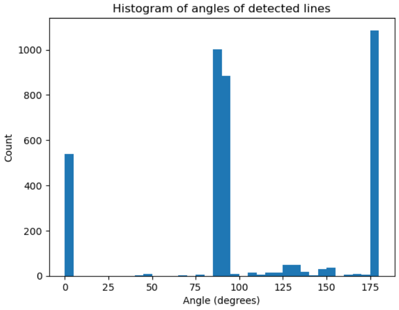

Figure 2. Road orientation bias in the original dataset.

IV. Models

Deep learning has revolutionized road segmentation by employing encoder-decoder architectures that balance accuracy and efficiency. We evaluate U-Net, LinkNet, and SegNet, each with unique strengths tailored for segmentation tasks.

A. U-Net Architecture

The U-Net, originally proposed for biomedical image segmentation [4], is a fully convolutional neural network designed for pixel-wise segmentation. It follows an encoder-decoder architecture: the encoder reduces spatial dimensions while capturing context through convolution and max pooling, and the decoder restores spatial resolution using up-convolutions.

A key feature of U-Net is its skip connections, which copy feature maps from the encoder to the corresponding decoder layers. This preserves fine-grained details and improves spatial recovery. The network outputs a segmentation map where each pixel is classified into a target class (e.g., road or background).

As shown in Figure 3, U-Net's multiscale feature extraction, skip connections, and symmetric design enable strong performance, particularly with limited training data.

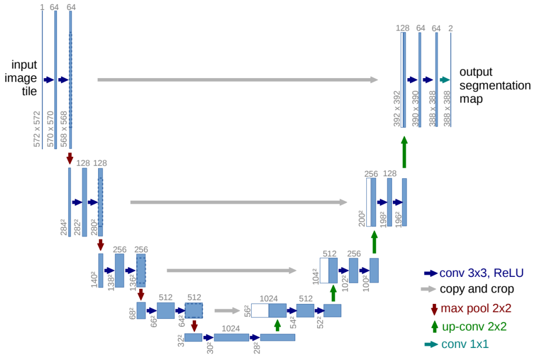

Figure 3. The U-Net architecture for segmentation tasks [4].

B. LinkNet and Its Variants

LinkNet is a lightweight CNN-based architecture for semantic segmentation [5], using an encoder-decoder framework with skip connections to retain spatial details. It employs a ResNet backbone for feature extraction, leveraging residual blocks to facilitate deeper networks by learning residual mappings instead of direct transformations. This design mitigates the vanishing gradient problem, enabling efficient multiscale feature extraction.

Variants Implemented: We tested three LinkNet variants:

- **LinkNet34:** Based on ResNet34, this variant is efficient and well-suited for real-time tasks.
- **NL-LinkNet:** Extends LinkNet34 with non-local blocks to capture long-range dependencies, improving segmentation in complex scenes.
- **D-LinkNet:** A notable extension, D-LinkNet [6], enhances LinkNet with dilated convolutions for a larger receptive field and a bottleneck decoder for precise upsampling, achieving strong results in the DeepGlobe 2018 Road Extraction Challenge.

C. SegNet Architecture

SegNet is an efficient encoder-decoder network designed for pixel-wise segmentation with low memory and computational requirements [7].
As shown in Figure 4, SegNet comprises:

- **Encoder:** Based on VGG16, it includes 13 convolutional layers with batch normalization, ReLU activation, and max-pooling (2×2, stride 2). To save memory, only max-pooling indices are stored instead of feature maps.
- **Decoder:** Each decoder layer upsamples using stored max-pooling indices, followed by convolution, batch normalization, and ReLU activation to generate dense feature maps.
- **Classification:** The final decoder outputs feature maps fed to a softmax classifier for multi-class tasks or a sigmoid activation for binary segmentation.

Originally designed for multi-class tasks, SegNet's memory-efficient architecture makes it ideal for binary road segmentation, enabling precise boundary delineation and pixel-level accuracy.

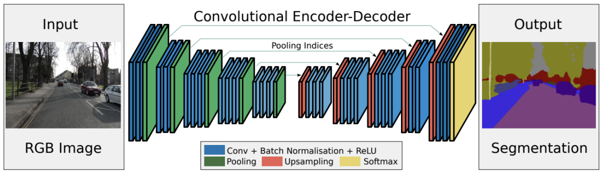

Figure 4. The SegNet architecture: a convolutional encoder-decoder network for semantic segmentation [7].

V. Model Training and Loss Functions

A. Training Framework

To evaluate the selected models (UNet, SegNet, and LinkNet variants), we implemented a flexible training pipeline with hyperparameter tuning and early stopping. Models were evaluated on a validation set using key metrics such as F1 score, Intersection over Union (IoU), and binary classification accuracy. The key training strategies were:

- **Batch Size:** Tested values of 8 and 16 to balance memory usage and efficiency.
- **Epochs:** Training was capped at 100 epochs with an early stopping criterion based on the validation F1 score, halting training after 5 consecutive epochs without improvement.
- **Learning Rate:** We experimented with learning rates between 1×10^{-5} and 1×10^{-3} while using a scheduler to reduce the rate by 5% every 5 epochs.
- **Threshold:** Predictions were binarized using a threshold of 0.25, classifying pixel values greater than 0.25 as roads (foreground) and the rest as background.

B. Loss Functions

Loss functions were categorized into three types:

- **Distribution-Based Loss Functions:** Include Binary Cross-Entropy (BCE) and Focal Loss, focusing on pixel-wise classification and emphasizing hard-to-predict pixels.
- **Region-Based Loss Functions:** Examples include Jaccard Loss and Dice Loss, which optimize spatial overlap. Squared Dice Loss improves gradient smoothness for better segmentation.
- **Compound Loss Functions:** These combine multiple objectives, such as BCE + Dice Loss or Combo Loss, to balance pixel accuracy and spatial consistency.

Based on [2] and our experiments, Binary Cross-Entropy (BCE) and Squared Dice Loss provided the most promising results, with BCE particularly effective for our best-performing configuration.

C. Evaluation and Early Stopping

Model performance was evaluated on the validation set using F1 score, Intersection over Union (IoU), and accuracy. Early stopping ensured training efficiency by terminating when no further improvements were observed.

VI. Experimental Results

We conducted multiple training runs with variations in batch size, learning rate, and loss functions. The table below presents the best results obtained for each model, evaluated using the AI Crowd platform. The reported F1 score and accuracy reflect the performance of the models on the test set.

Table I
Best Results for SegNet, UNet, NL_LinkNet and DLinkNet on the Test Set (AI Crowd Evaluation)

Model	lr	Batch Size	Loss Function	F1 Sc.	Acc.
SegNet	3e-4	16	BCE	0.87	0.94
UNet	4e-4	16	Squared Dice	0.83	0.91
NL_LinkNet	3e-4	16	Squared Dice	0.85	0.92
DLinkNet	3e-4	16	Squared Dice	0.859	0.915

The SegNet model achieved the best overall performance, with an F1 score of 0.87 and accuracy of 0.94, followed by NL_LinkNet, DLinkNet and UNet.

A. Training and Validation Loss Analysis

To better understand the training behavior, we plotted the evolution of training and validation losses for three of the models. Figure 8 shows a comparison of loss curves for SegNet, UNet, and NL_LinkNet.
The SegNet model demonstrates steady convergence after around 15 epochs, with a small gap between training and validation losses, indicating minimal overfitting. This reflects the effectiveness of BCE loss in binary segmentation and the slower initial decline due to the model's larger architecture.

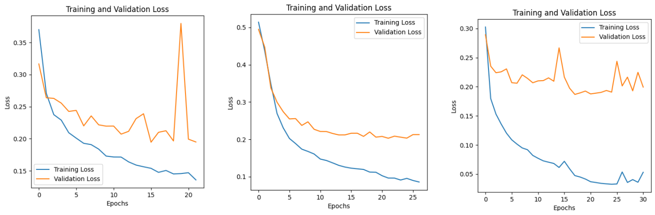

Figure 5. *
(a) SegNet

Figure 6. *
(b) UNet

Figure 7. *
(c) NL_LinkNet

Figure 8. Training and validation loss comparison for SegNet, UNet, and NL_LinkNet. These plots correspond to the best performances achieved for each model, with hyperparameters such as batch size and learning rate detailed in Table I.

For UNet, training loss decreases faster than validation loss, particularly early on, suggesting mild overfitting likely due to sensitivity to Squared Dice Loss. While losses stabilize after 10 epochs, the persistent gap indicates potential benefits from additional regularization like dropout.

The NL_LinkNet converges fastest, leveraging its pre-trained ResNet34 encoder for rapid learning. Both training and validation losses align closely, highlighting strong generalization and the advantage of pre-trained encoders for segmentation tasks.

These results emphasize the influence of architectural design and loss functions on convergence and generalization in binary road segmentation.

B. Test Set Predictions and Analysis

To evaluate model performance on the test set, we compared predicted segmentation masks for a sample image (Fig. 9). This image, labeled *Original Image*, showcases urban roads with distinct boundaries, providing a clear basis for assessing segmentation quality.

Key Observations:

- **SegNet:** Demonstrated the best F1 score (0.87) and accuracy (0.94), excelling at comprehensive road detection but prone to over-segmentation and less precise boundaries.
- **UNet:** Delivered balanced performance with effective road region detection but struggled with fine boundary details, especially at intersections.
- **LinkNet:** Produced the sharpest and cleanest segmentation masks, leveraging its pre-trained encoder. However, it occasionally missed smaller road segments, slightly reducing recall and F1 score.

These results reveal a trade-off between visual quality and performance metrics. SegNet's architecture prioritizes recall, inflating the F1 score by accurately detecting road regions at the expense of boundary precision. Conversely, LinkNet achieves sharper masks with better boundary delineation but compromises recall slightly by missing smaller roads. UNet offers a balance but struggles in complex intersections.

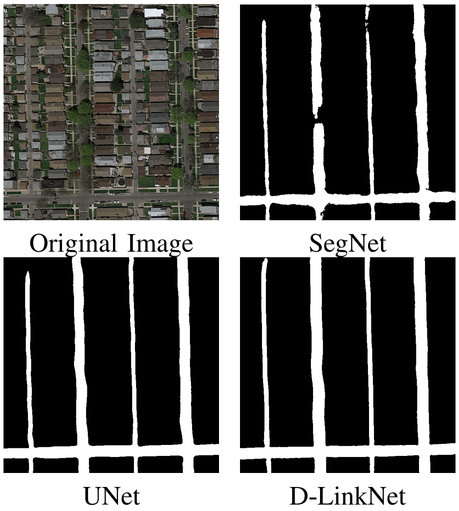

Figure 9. Comparison of predicted segmentation masks on a sample test image. D-LinkNet predictions are shown, as NL-LinkNet results were visually similar.

For applications requiring detailed boundary precision, LinkNet is a suitable choice. However, for tasks emphasizing complete road coverage, such as rural mapping or disaster response, SegNet's higher recall and F1 score make it more appropriate.

VII. Conclusion

This study compared U-Net, SegNet, and LinkNet variants for road segmentation in high-resolution imagery. SegNet achieved the best metrics, with an F1 score of 0.87 and accuracy of 0.94, though at the cost of sharper boundary precision. Incorporating a diverse external dataset and effective data augmentation improved model generalization. Binary Cross-Entropy, combined with SegNet, proved particularly effective.

The trade-offs between metrics and visual quality underline the need to align model selection with specific application goals. While SegNet excels in road coverage, LinkNet's sharper masks are better for precise delineation. These insights provide a foundation for robust road segmentation pipelines across navigation, urban planning, and emergency response.

Ethical risks

We identified a welfare risk related to the potential bias in road segmentation models, which arises primarily from a lack of data in our given dataset, which at the beginning contained only 150 images and also from an imbalance in the training data. This limited dataset fails to adequately capture the diversity of environments, such as rural or less developed areas, in comparison to more urban regions. But also the diversity of road structures. This imbalance can lead to models that are less accurate in underrepresented areas.

There are two types of stakeholders, direct stakeholders and indirect stakeholders. The direct Stakeholders include urban planners, transportation agencies, and users of navigation systems that rely on accurate segmentation outputs for decision-making. The indirect Stakeholders include communities in underrepresented regions, where the model may struggle to detect road networks correctly, potentially excluding these areas from critical applications, such as emergency response and urban planning.

The negative impact can be significant. In areas underrepresented in the dataset, inaccurate segmentation could result in poor road mapping, which would negatively impact navigation and infrastructure development.

For the risk evaluation, the severity of this risk is high, as errors in road segmentation can affect transportation planning, navigation safety, and access to essential services and resources. In terms of likelihood, the likelihood of this risk occurring is medium, because data imbalance is a common issue in remote sensing datasets, it can be solved with proper strategies.

The risk can be evaluated using quantitative metrics, such as accuracy and the F1 score, to evaluate performance. A scenario where accuracy is high but F1 scores are low may indicate issues related to dataset imbalance. Additionally, segmentation outputs should be qualitatively analyzed to identify discrepancies, particularly in underrepresented areas. For example, road segmentation images generated during predictions can be examined to determine if they are less precise in underrepresented regions. We can also examined the training and testing datasets at the beginning to evaluate their diversity in terms of geographic and environmental conditions.

To address the identified risks, several mitigation strategies were implemented. Data augmentation was employed by incorporating additional datasets that represent diverse geographic regions worldwide. Rotations and adjustments in brightness were applied to reduce biases related to orientation and lighting conditions, along with the introduction of noise and random transformations to enhance model robustness. The specific modifications applied are detailed in the data augmentation and preparation part. Additionally, we experimented with customizing the loss function to improve the model's handling of edge cases and smaller road features. Alternative loss functions, such as Squared Dice Loss, were tested to enhance segmentation performance in challenging scenarios.

However, some challenges were encountered during the mitigation process. Preprocessing images, such as resizing and applying transformations like rotations, required careful attention to ensure the augmented data aligned with the quality and characteristics of the original dataset, including padding adjustments. Additionally, identifying and integrating an external dataset that was compatible with the original dataset required additional effort to maintain consistency and quality.

References

[1] A. Abdollahi, B. Pradhan, N. Shukla, S. Chakraborty, and A. Alamri, "Deep learning approaches applied to remote sensing datasets for road extraction: A state-of-the-art review," *Remote Sensing*, vol. 12, no. 9, 2020. [Online]. Available: https://www.mdpi.com/2072-4292/12/9/1444

[2] H. Xu, H. He, Y. Zhang, L. Ma, and J. Li, "A comparative study of loss functions for road segmentation in remotely sensed road datasets," *International Journal of Applied Earth Observation and Geoinformation*, vol. 116, p. 103159, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1569843222003478

[3] P. Kaiser, J. D. Wegner, A. Lucchi, M. Jaggi, T. Hofmann, and K. Schindler, "Learning aerial image segmentation from online maps," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 11, p. 6054–6068, Nov. 2017. [Online]. Available: http://dx.doi.org/10.1109/TGRS.2017.2719738

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015. [Online]. Available: https://arxiv.org/abs/1505.04597

[5] A. Chaurasia and E. Culurciello, "Linknet: Exploiting encoder representations for efficient semantic segmentation," in *2017 IEEE Visual Communications and Image Processing (VCIP)*, 2017, pp. 1–4.

[6] L. Zhou, C. Zhang, and M. Wu, "D-linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[7] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," 2016. [Online]. Available: https://arxiv.org/abs/1511.00561