Road Segmentation in Aerial Imagery using U-Net Ensembles

Hervé Sérandour, Jean Marc P. Fata, Lina Obaid
EPFL, Switzerland

***Abstract*—Road segmentation from aerial imagery is a fundamental task in remote sensing, with direct applications in automated mapping and urban planning. In this project, we present a comparative analysis of classical machine learning and state-of-the-art deep learning approaches for binary road detection. Three methodologies are implemented and evaluated: (1) a Random Forest baseline based on patch-level feature extraction, where 16×16 pixel blocks are classified using aggregated spectral, textural, and edge-based statistics; (2) a custom Convolutional Neural Network (CNN) trained end-to-end using binary cross-entropy with logits to address class imbalance; and (3) a deep learning ensemble combining a U-Net++ with a ResNet34 encoder and a standard U-Net with an EfficientNet-B3 encoder. To ensure robust evaluation and mitigate overfitting on the limited dataset, the ensemble model is validated using 5-fold cross-validation with deep model ensembling. Experimental results reveal a clear performance hierarchy: while the patch-based Random Forest provides a computationally efficient baseline, it produces blocky artifacts and lacks fine-grained boundary precision. In contrast, the U-Net ensemble achieves superior performance, effectively capturing global road topology and preserving geometric continuity. These findings highlight the necessity of deep semantic segmentation architectures for accurately modeling the complex textures present in high-resolution aerial imagery.**

I. Introduction

Road segmentation from aerial imagery is a core problem in computer vision and remote sensing, with applications in automated mapping, urban planning, and autonomous navigation systems. The objective is to perform pixel-wise binary classification, assigning each pixel either to the road class (1) or to the background class (0).

This task is particularly challenging due to variations in road appearance, inconsistent illumination conditions, shadows, and occlusions caused by buildings or vegetation. To address these challenges, supervised learning methods are required to extract discriminative representations from labeled data.

In this project, we study the trade-offs between classical machine learning techniques and modern deep learning architectures for road segmentation. We implement and evaluate the following three methodologies:

- **Random Forest (Patch-Based Baseline):** A classical machine learning model trained on handcrafted features extracted from 16×16 pixel patches. To reduce the loss of spatial resolution inherent in block-wise classification, we employ overlapping patch aggregation to smooth predictions.
- **Custom Convolutional Neural Network (CNN):** A lightweight, custom-designed CNN trained end-to-end. This model serves as an intermediate baseline to assess the benefits of local texture learning, while omitting skip connections, attention mechanisms, and multi-scale feature fusion commonly used in deeper semantic segmentation models.
- **Deep Learning Ensemble (U-Net++ & U-Net):** A robust ensemble combining a U-Net++ with a **ResNet34** encoder and a standard U-Net with an **EfficientNet-B3** encoder. To ensure reliable performance estimation and improved generalization on the limited dataset, this approach incorporates 5-fold cross-validation and deep model ensembling, and represents our best and most advanced solution.

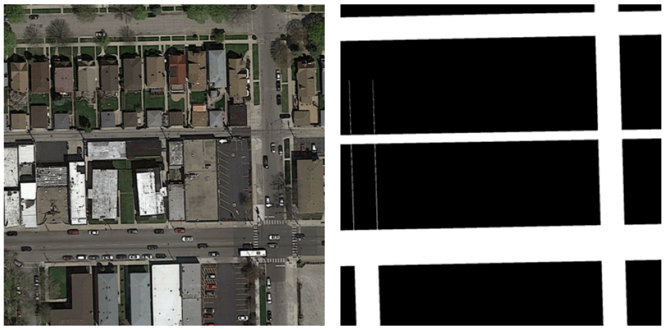

Fig. 1. Example aerial image from the dataset (Left) and its corresponding binary segmentation mask (Right).

The performance of these models is compared using segmentation-relevant metrics such as the F1 score and accuracy, with a focus on understanding their respective strengths, limitations, and computational trade-offs. This report details the methodology, implementation, and experimental findings of this comparative study.

II. Dataset and Preprocessing

The dataset used in this project consists of high-resolution aerial images, making it well-suited for road segmentation tasks. The original dataset is divided as follows:

- **Training set:** 100 images, each paired with a corresponding binary ground truth mask (see Figure 1).
- **Test set:** 50 images, provided without labels for blind evaluation.

Each image has a resolution of 400×400 pixels with 3 RGB color channels. These images capture aerial views of roads and surrounding landscapes.

A. *Validation Strategy*

Given the limited dataset size (100 images), a single static train/validation split could lead to biased performance estimates. Therefore, we employed two distinct validation strategies depending on the model:

- **For Baseline Models (RF & CNN):** A static 80/20 random split (80 images for training, 20 for validation) was used to establish initial performance benchmarks and tune hyperparameters.
- **For the U-Net Ensemble:** We implemented **5-Fold Cross-Validation**. The entire dataset was divided into 5 folds; in each iteration, the model was trained on 80 images and validated on the remaining 20. This ensures that every image in the dataset is used for validation exactly once, providing a robust estimate of the model's generalization error.

B. *Preprocessing and Augmentation*

Different models required distinct data preparation strategies to match their architectural requirements:

- **Random Forest (Feature Engineering):** Images were normalized to the range $[0, 1]$. To enable patch-based classification, we extracted 16×16 pixel patches. Crucially, we computed binary labels based on a **fixed road pixel density threshold** (0.25) to ensure robust ground truth generation.
- **Deep Learning Models:** Different strategies were applied based on architectural requirements. For the U-Net Ensemble, images were resized to 384×384 pixels (multiples of 32) and normalized using standard ImageNet mean ($\mu = [0.485, 0.456, 0.406]$) and standard deviation ($\sigma = [0.229, 0.224, 0.225]$) [1] [2]. In contrast, the custom CNN operated on the full native resolution (400×400 pixels) using dataset-specific normalization.
- **Data Augmentation:** To prevent overfitting on the deep learning models, we applied a heavy augmentation pipeline during training. This included horizontal and vertical flips, random 90-degree rotations, and Shift-Scale-Rotate transformations to enforce rotational invariance.

a) Random Forest Specifics: For the training phase, images were divided into **non-overlapping patches** to prevent data leakage between training samples. To address the significant class imbalance without losing training data, the model was trained using **cost-sensitive learning** (`class_weight='balanced'`), which assigns higher penalties to misclassifying the minority road class. However, during the inference phase, we extracted **overlapping patches** (stride 8) to aggregate predictions. The final binary mask was generated using a probability threshold optimized via grid search on the validation set, rather than a default 0.5 cutoff, to maximize the F1 score.

b) CNN & U-Net Specifics.: For the Deep Learning models, we utilized full-resolution input (resized to 384×384). Unlike the Random Forest, which relies on handcrafted features, these models learned features end-to-end. We addressed class imbalance in the custom CNN using a **binary cross-entropy with logits**, while the U-Net utilized a hybrid objective function ($L = 0.5 L_{Dice} + 0.5 L_{BCE}$). This combination balances the pixel-wise accuracy of Binary Cross Entropy with the region-based overlap optimization of the Dice loss, which is critical for handling the geometric continuity of roads [3].

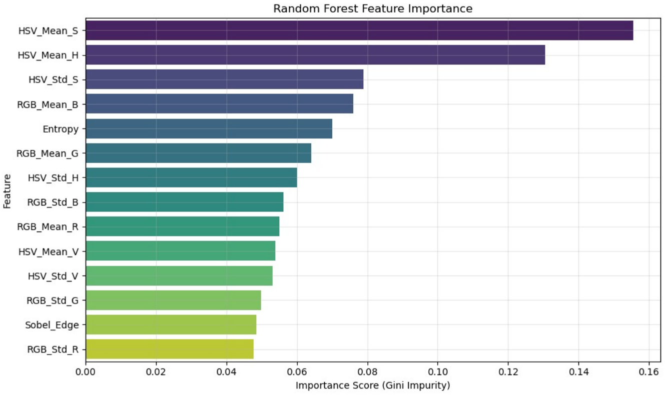

Fig. 2. Importance of each of the 14 Features

III. Models

In this project, we explored a progression of architectures, incrementally adopting more advanced techniques from classical machine learning to deep ensembles.

Random Forest Model (Baseline)

The Random Forest baseline predicts labels at the patch level (16×16 pixels) rather than predicting individual pixels.

Feature Extraction: We extract a 14-dimensional feature vector per patch, combining **spectral statistics** (RGB/HSV mean and std) with **texture and structure** (Shannon Entropy and Sobel edge density). We specifically integrated HSV to decouple chromaticity from intensity, improving robustness to shadows. Feature importance analysis (Figure 2) confirms that these spectral and textural features are critical for distinguishing road surfaces.

Inference: While training relied on a fixed density threshold (0.25) to generate binary ground truth labels, inference required a dynamic approach. We optimized the decision boundary by performing a grid search on the validation set, selecting the probability threshold that maximized the F1 score. Predictions were generated using overlapping patches (stride 8, 50% overlap) and aggregated via spatial voting.

Convolutional Neural Network (CNN)

We implemented a custom convolutional neural network as a deep learning baseline for the road segmentation task.

Architecture: The model follows a lightweight fully convolutional encoder–decoder design, composed of convolutional layers with ReLU activations and Batch Normalization. Spatial resolution is reduced using max-pooling operations and later restored via bilinear upsampling. Unlike the Random Forest baseline, which relies on disjoint image patches, the CNN processes the input image as a continuous signal, preserving local spatial relationships throughout the network.

Training Strategy: Input images were converted to tensors and standardized using dataset-specific mean and standard deviation. The network was optimized using AdamW, which has been shown to provide improved generalization compared to standard SGD. To address the significant class imbalance inherent to road segmentation—where road pixels are rare relative to background—we employed a weighted Binary Cross-Entropy loss with logits (`BCEWithLogitsLoss`). A positive class weight of 2.0 was empirically selected to penalize false negatives more heavily, thereby improving recall on thin road structures without introducing excessive noise, a common trade-off in imbalanced segmentation tasks.

Deep Learning Ensemble (U-Net++)

Our final and most performant approach is a robust ensemble designed to maximize segmentation accuracy. We leveraged transfer learning by using encoders pre-trained on ImageNet to extract high-level semantic features.

Architectures: The ensemble aggregates two distinct variants of the U-Net architecture to ensure diversity:

- **Model A (U-Net++ with ResNet34):** The nested skip pathways of U-Net++ reduce the semantic gap between the encoder and decoder [4], improving the segmentation of fine details like narrow roads. We utilized a ResNet34 backbone for robust feature extraction.
- **Model B (Standard U-Net with EfficientNet-B3):** Chosen for its high parameter efficiency and ability to capture broader contextual features.

Both models were minimized using a hybrid objective ($L = 0.5L_{BCE} + 0.5L_{Dice}$) to balance pixel-wise accuracy with the structural overlap of road regions. To ensure robust generalization, we employed **5-Fold Cross-Validation**. During inference, we applied **deep model ensembling**, averaging predictions from all 10 trained models (2 architectures × 5 folds) to produce the final segmentation mask.

IV. Training and Implementation Details

To ensure reproducibility and consistency, the entire pipeline was implemented using the `PyTorch` framework [5] with a fixed random seed (`SEED=42`). Deep learning architectures and pre-trained weights were sourced from the `segmentation-models-pytorch` library [6], while robust data augmentation was performed using `Albumentations` [7]. Classical machine learning models and metrics were implemented via `Scikit-Learn` [8]. Training was executed on a Google Colab environment utilizing an NVIDIA T4 GPU to accelerate deep learning computations, while the Random Forest training was executed on CPU.

Below are the detailed training protocols and hyperparameter settings used for each architecture.

- **Optimization (Deep Learning):**
  - *CNN:* Optimized using **AdamW** with a fixed learning rate of 3×10^{-4} and a weight decay of 1×10^{-3} to prevent overfitting.
  - *U-Net Ensemble:* Optimized using **AdamW** with an initial learning rate of 1×10^{-4}. We employed a **Cosine Annealing Warm Restarts** scheduler ($T_0 = 10, T_{mult} = 2$), which periodically resets the learning rate to escape local minima and find more robust solutions, allowing the model to fine-tune its weights in later epochs.
- **Loss Functions:**
  - *CNN:* A **Weighted Binary Cross Entropy with Logits** loss (‘pos_weight=2.0‘) was used to explicitly penalize false negatives (missed roads) and address class imbalance.
  - *U-Net Ensemble:* A hybrid objective function was minimized: $L = 0.5L_{Dice} + 0.5L_{BCE}$, balancing pixel-wise accuracy (BCE) with the geometric overlap of road regions (Dice).
- **Early Stopping:** Training was monitored using the validation F1 score. For the CNN, training was terminated if performance did not improve for 5 consecutive epochs (‘patience=5‘). For the U-Net, the model weights achieving the **highest validation F1 score** were automatically retained.
- **Threshold Selection (Inference):**
  - *Random Forest:* We performed a **Grid Search** on the validation set to determine the optimal probability threshold.
  - *Deep Learning Models:* We utilized a standard classification threshold of 0.5, as the sigmoid activations from the ensemble (averaged via deep model ensembling) provided well-calibrated probability maps.

V. Results and Analysis

A. Quantitative Performance

Given the severe class imbalance ($\approx$ 10% road pixels), we prioritized the **F1 Score** to evaluate model performance. Table I summarizes the results. The **Random Forest** baseline (Test F1: 0.670) struggled with spatial continuity. The **Custom CNN** significantly improved generalization (Test F1: 0.776). The **U-Net Ensemble**, validated via 5-fold cross-validation, achieved superior results across all metrics, attaining a Test F1 of **0.895** and a Validation Accuracy of **94.7%**, confirming the robustness of the deep ensembling strategy.

Model	**Val. F1**	**Val. Acc.**	**Test F1 (AICrowd)**
Random Forest	0.661	77.61%	0.670
Custom CNN	0.690	89.00%	0.776
U-Net Ensemble	**0.897**	**94.70%**	**0.895**

TABLE I
Performance comparison. The U-Net Ensemble demonstrates superior generalization on the blind test set.

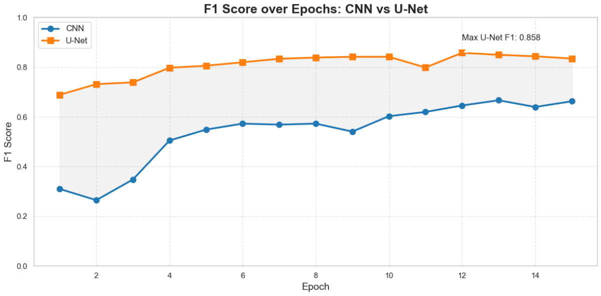

Fig. 3. Validation F1 convergence. The Ensemble (Orange) shows faster and more stable learning than the single CNN (Blue).

B. Qualitative Analysis

The visual results (Figure 4) corroborate the quantitative metrics and highlight the architectural trade-offs:

- **Random Forest:** Despite using overlapping patch aggregation (stride 8) to smooth predictions, the output exhibits characteristic blocky artifacts. It successfully detects broad road segments but struggles with curvature and fine boundaries due to the fixed 16 × 16 patch resolution.
- **Deep Learning Models:** The CNN and U-Net models produce significantly smoother and more coherent segmentations. The U-Net Ensemble, in particular, effectively captures global road topology and preserves geometric continuity, resolving the fragmentation issues seen in the baseline.

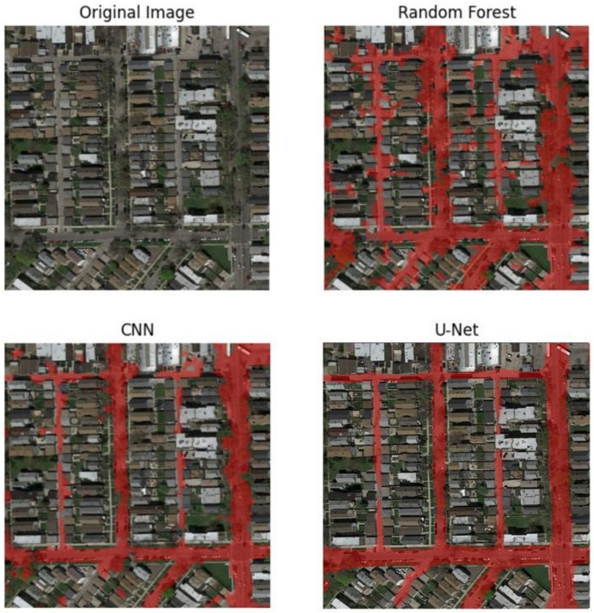

Fig. 4. Qualitative comparison of road segmentation predictions produced by the different models on a test image. Red regions indicate predicted road areas

VI. Ethical Risks

An important ethical risk identified in this project relates to dataset bias and its downstream effects on public decision-making. The key aspects of this risk are summarized as follows:

- **Source of risk:** The dataset is predominantly composed of urban satellite images, leading to limited representation of rural and low-income regions.
- **Potential harm:** Models trained on such data may produce missing or fragmented road predictions in underrepresented areas, distorting assessments of connectivity and infrastructure coverage.
- **Affected stakeholders:** Local governments and territorial planners, who depend on accurate road maps for infrastructure planning and equitable resource allocation.
- **Severity and likelihood:** The severity is high, as segmentation errors can influence long-term infrastructure decisions. The likelihood is reinforced by documented urban–rural performance gaps in remote sensing segmentation systems.
- **Evaluation:** We performed a qualitative visual audit of urban and rural image tiles and reviewed relevant remote sensing literature and the *Fairness* section of the Digital Ethics Canvas (2024). No quantitative metrics were used, as the focus was on dataset representativeness.[9]
- **Mitigation:** The risk could not be mitigated due to the limited size and scope of the dataset, which prevented rebalancing or geographic expansion.

This analysis highlights the importance of diverse and representative datasets for ethical deployment of road segmentation systems.

VII. Conclusion

In this project, we explored multiple approaches to the problem of road segmentation from aerial imagery, ranging from classical machine learning techniques to deep learning-based semantic segmentation models. By evaluating a Random Forest model, a convolutional neural network, and a custom U-Net architecture under a unified training and evaluation framework, we were able to assess the impact of model complexity and spatial reasoning on segmentation performance.

The results demonstrate that deep learning models, particularly CNN-based approaches, generally provide higher pixel-wise accuracy and more coherent segmentation outputs. However, the Random Forest model achieved competitive F1 scores, illustrating that handcrafted features combined with overlapping patch aggregation can still extract meaningful structural information from aerial images. This highlights the value of classical methods as strong baselines, especially in scenarios with limited computational resources or training data.

Overall, this work emphasizes the importance of consistent preprocessing, careful threshold selection, and appropriate evaluation metrics when comparing segmentation models. Future work could focus on improving the U-Net implementation, incorporating more advanced data augmentation strategies, and evaluating generalization performance on more diverse and geographically representative datasets to further enhance robustness and ethical deployment. The author thanks Christian Sigg for his careful reading and helpful suggestions.

REFERENCES

[1] PyTorch Team, "Torchvision models," https://docs.pytorch.org/vision/0.12/models.html, 2022, accessed: 2025-01-18.

[2] Wikipedia contributors, "Imagenet," https://en.wikipedia.org/wiki/ImageNet, 2025, accessed: 15 December 2025.

[3] V. Rajput, "Robustness of different loss functions and their impact on networks learning capability," *arXiv:2110.08322*, 2021, https://arxiv.org/abs/2110.08322 (preprint).

[4] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," *arXiv preprint arXiv:1807.10165*, 2018. [Online]. Available: https://arxiv.org/abs/1807.10165

[5] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019.

[6] P. Yakubovskiy, "Segmentation models pytorch," https://github.com/qubvel/segmentation_models.pytorch, 2020.

[7] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parakhin, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, 2020. [Online]. Available: https://www.mdpi.com/2078-2489/11/2/125

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[9] M. Zhang and R. Chunara, "Mitigating urban-rural disparities in contrastive representation learning with satellite imagery," *arXiv preprint arXiv:2211.08672*, 2022.