
High-Resolution Road Extraction from Remote Sensing Imagery Using Deep Learning: A Comparative Study

Wei Jiang¹, Shamsul Arrieya Bin Ariffin²

¹ City University of Malaysia, Malaysia,

²Faculty of Computing and Meta-Technology, Sultan Idris Education University, Malaysia

Abstract

Introduction: *Accurate road extraction from high-resolution remote sensing imagery (HRRSI) is vital for urban planning, transportation management, and emergency response. Traditional rule-based or pixel-level algorithms often fail in complex environments due to shadow interference, occlusion, and spectral similarity between roads and other objects. This study aims to evaluate a deep learning framework based on convolutional neural networks (CNNs) for automatic road extraction and to systematically analyze the effects of loss functions, optimizers, and spatial resolution on model performance.*

Methodology: *A dual-source dataset comprising satellite imagery (0.5 m) and unmanned aerial vehicle (UAV) imagery (0.05 m) from Zhanggong District, Jiangxi Province, China, was used. A modified U-Net model was trained with two loss functions (Binary Cross-Entropy and BCE + Dice) and four optimizers (SGD, Adam, AdamW, RMSprop). Quantitative metrics—including Pixel Accuracy (PA), Intersection over Union (IoU), Precision, Recall, and F1-score—were employed for evaluation.*

Results and Discussion: *The combination of BCE loss and RMSprop optimizer achieved the highest and most stable performance, with IoU and F1-scores of 80.8 % and 89.38 % on satellite imagery and 84.53 % and 91.62 % on UAV imagery. UAV data provided better boundary continuity, whereas satellite imagery maintained higher global consistency. Comparative analyses demonstrated that CNN-based segmentation significantly outperformed traditional and machine-learning baselines.*

Conclusion and Recommendations: *Deep learning substantially enhances the accuracy and robustness of road extraction in HRRSI. Future research should focus on integrating multi-resolution imagery, attention mechanisms, and topology-aware optimization to improve continuity and scalability for operational mapping applications.*

Keywords: *Remote sensing imagery; Road extraction; Deep learning; Convolutional neural network (CNN); Optimization strategy; High-resolution imagery; UAV*

1. Introduction

The extraction of road networks from high-resolution remote sensing imagery (HRRSI) has long been a central topic in the fields of photogrammetry, geoinformatics, and intelligent transportation. As the backbone of national infrastructure, roads support economic development, facilitate spatial connectivity, and serve as indispensable data layers in urban planning, emergency management, and digital-twin city modelling. The increasing availability of sub-meter satellite imagery and high-resolution unmanned aerial vehicle (UAV) data has significantly improved our capability to observe, model, and manage transportation systems. Yet, despite these technological advances, the automated and accurate delineation of road networks remains a challenging task due to the complexity of imaging conditions, heterogeneous landscapes, and the intrinsic spectral ambiguity between roads and surrounding ground objects such as rooftops, parking lots, and bare soil (Wang et al., 2016).

1.1 Background and significance

Traditional approaches to road extraction primarily rely on handcrafted features and heuristic rules derived from spectral, geometric, and topological attributes (Gaetano et al., 2011). Early techniques utilized edge detectors such as the Canny or Sobel operators to identify linear features, followed by morphological filtering or Hough-transform-based fitting to reconstruct continuous road segments. Although effective for well-defined asphalt roads under uniform illumination, these methods tend to fail when encountering shadowed areas, occlusions, or heterogeneous surfaces. Moreover, the reliance on manually tuned thresholds limits adaptability across varying spatial resolutions and geographic contexts (Das et al., 2011).

The rapid development of artificial intelligence, particularly deep learning (DL), has opened new avenues for automated feature extraction from complex imagery. By mimicking the hierarchical representation learning of human vision, convolutional neural networks (CNNs) have demonstrated superior performance in numerous vision tasks, including object detection, semantic segmentation, and scene understanding (Ronneberger et al., 2015). Their ability to learn multiscale spatial-spectral features directly from raw data without human-defined descriptors provides a theoretical foundation for robust and generalizable road extraction models.

From a practical perspective, accurate and timely road information derived from HRRSI is critical for a broad range of applications. In urban planning, extracted roads provide geometric constraints for parcel segmentation, land-use classification, and 3D city modelling. In disaster management, near-real-time road extraction enables the identification of blocked or damaged routes for emergency logistics and rescue planning. In intelligent transportation, updated road networks form the backbone of high-definition maps used in autonomous navigation and traffic simulation. The integration of deep-learning-based road mapping within geographic information systems (GIS) thus represents an essential step toward fully automated spatial data infrastructures.

1.2 Challenges in high-resolution road extraction

Despite remarkable progress, several intrinsic challenges persist.

First, spectral similarity and background interference frequently cause misclassification. The spectral signatures of roads can overlap with those of buildings, concrete yards, and dry riverbeds, particularly in panchromatic or RGB imagery lacking multispectral depth.

Second, spatial heterogeneity of road surfaces—ranging from wide multilane highways to narrow rural paths—introduces large variations in texture, width, and brightness. Standard convolution kernels often fail to simultaneously capture both broad and fine-scale features, leading to fragmented predictions.

Third, shadow and occlusion problems remain significant. Shadows cast by trees or tall structures can dramatically alter pixel intensities, while vehicles and temporary obstacles interrupt linear continuity.

Finally, data annotation and generalization constitute practical bottlenecks. High-quality pixel-level labels are labor-intensive to produce, and models trained on limited areas frequently underperform when transferred to new environments or sensors. These issues highlight the need for a systematic evaluation of loss functions, optimization strategies, and data resolution effects within deep-learning road extraction frameworks.

1.3 Deep learning for remote sensing interpretation

Deep learning has revolutionized remote sensing analysis by enabling end-to-end feature learning and large-scale semantic segmentation. Architectures such as U-Net (Ronneberger et al., 2015), DeepLab v3+ (Chen et al., 2018), and D-LinkNet (Zhou et al., 2018) have achieved state-of-the-art accuracy in land-cover classification and object extraction. In road extraction, CNN-based models learn hierarchical representations that integrate local edge cues with global contextual dependencies, improving the detection of narrow, elongated structures.

Nevertheless, model design choices—including loss formulation, optimizer selection, and data resolution—strongly influence performance. The loss function determines how prediction errors are penalized and thus directly affects convergence and boundary precision. The optimizer governs parameter updates and learning dynamics, balancing convergence speed against generalization. While numerous studies have focused on architectural innovation, relatively few have conducted systematic comparative analyses of these training components under consistent experimental conditions.

1.4 Research motivation and objectives

Recognizing the gaps outlined above, this study aims to provide a comprehensive comparative analysis of deep-learning-based road extraction methods for HRRSI. The research emphasizes two intertwined dimensions: algorithmic optimization and data-source variability. Specifically, we investigate how different combinations of loss functions and optimizers affect model performance across satellite and UAV imagery with distinct spatial resolutions.

The main objectives are as follows:

- A. To construct a robust experimental framework for CNN-based road extraction using dual-source high-resolution datasets.
- B. To evaluate and compare the effects of two widely used loss functions—Binary

Cross-Entropy (BCE) and BCE+Dice—on model convergence, segmentation accuracy, and edge preservation.

C. To analyze the performance of four optimizers—SGD, Adam, AdamW, and RMSprop—under identical network and hyperparameter settings, identifying the most efficient training configuration.

D. To examine the influence of spatial resolution (0.5 m satellite vs 0.05 m UAV imagery) on extraction accuracy, continuity, and robustness in various environmental contexts.

E. To summarize best practices and provide methodological recommendations for future deep-learning road extraction studies in terms of data preparation, model selection, and performance evaluation.

2. Literature review

2.1 Evolution of road extraction from remote sensing imagery

The extraction of road networks from remotely sensed imagery has evolved through three distinct methodological eras: the traditional image processing era, the machine-learning era, and the deep-learning era. Each period reflects the technological advancement of sensors, data availability, and computational capability. The progression from heuristic pixel-based models to data-driven neural architectures represents a fundamental paradigm shift in how spatial information is interpreted and utilized.

2.1.1 Traditional image processing approaches

Early studies dating back to the 1970s and 1980s primarily relied on pixel-based statistical and geometric analysis. In these methods, the spectral characteristics of asphalt and concrete surfaces—usually with low reflectance in the visible spectrum—were used to distinguish roads from vegetation and built-up areas. For instance, edge detectors such as Canny and Sobel were applied to emphasize linear structures, followed by thresholding and morphological filtering to suppress noise (Gaetano, Zerubia, & Scarpa, 2011). Hough-transform-based algorithms were then used to fit continuous lines across detected segments, producing approximated road axes.

Although these early approaches were computationally efficient, they suffered from severe limitations in heterogeneous environments. Roads in urban scenes often share similar reflectance values with rooftops or parking lots, while in rural landscapes, dirt roads exhibit spectral confusion with bare soil. In addition, these algorithms were sensitive to illumination and viewing angles, leading to inconsistent performance across datasets (Senthilnath, Rajeshwari, & Omkar, 2009).

To mitigate these issues, researchers introduced texture-based descriptors and spatial context models. Das, Mirnalinee, and Varghese (2011) designed a multistage framework combining morphological filtering and texture progressive analysis (TPA), which segmented roads based on structural continuity. However, these improvements remained rule-dependent and lacked adaptability. As remote sensing imagery advanced from medium to high spatial resolution, the inadequacy of handcrafted thresholding became more evident.

2.1.2 Semi-automatic and model-based extraction

With the advent of Geographic Object-Based Image Analysis (GEOBIA) and segmentation theory, attention shifted toward object-level processing. Researchers attempted to exploit geometric primitives—such as rectangularity, elongation, and connectivity—to model roads more explicitly. For example, Shanmugam and Kaliaperumal (2016) proposed a dynamic color-thresholding method that incorporated a “junction-aware” water-flow model to improve connectivity between road segments. Although it enhanced network reconstruction, the reliance on pre-defined parameters constrained the algorithm’s transferability.

Model-based approaches, such as active contour models (Snakes) and level-set evolution, introduced energy-minimization principles to enforce spatial smoothness. Liu and Lin (1996) applied Snake models to semi-automatically delineate roads from aerial photographs, while subsequent work integrated genetic algorithms to optimize control points (Wu, 2005). These efforts paved the way for automatic extraction by embedding geometric constraints into the segmentation process. Nevertheless, they still required significant manual intervention for initialization, and their convergence depended heavily on parameter tuning.

2.1.3 Transition to machine learning

During the early 2000s, the proliferation of multispectral and high-resolution sensors generated a surge of feature-rich imagery, providing a fertile ground for machine-learning algorithms. Supervised classifiers such as Support Vector Machines (SVM), Random Forests (RF), and AdaBoost became popular alternatives to rule-based models (Das et al., 2011; Zhu et al., 2016). These methods utilized spectral indices (e.g., Normalized Difference Road Index), texture descriptors (e.g., GLCM), and geometric features (e.g., length-width ratios) as inputs for training discriminative models.

Although machine learning improved classification robustness and reduced manual tuning, its effectiveness remained bounded by the quality of manually engineered features. Moreover, the classifiers treated pixels or segments independently, lacking the contextual awareness needed to maintain network continuity. As a result, extracted road masks were often fragmented and disconnected.

Object-oriented classification frameworks further advanced the field by integrating region adjacency and topological rules. Approaches based on Markov Random Fields (MRF) and Conditional Random Fields (CRF) modeled contextual dependencies between neighboring segments. For example, Zhu, Song, and Dai (2016) proposed an MRF-regularized SVM classifier that enhanced consistency across urban scenes. However, these probabilistic graphical models entailed high computational costs and remained sensitive to initial segmentation accuracy. The need for methods capable of learning hierarchical spatial semantics thus became increasingly apparent.

2.2 Emergence of deep learning in remote sensing road extraction

2.2.1 Rise of convolutional neural networks (CNNs)

The breakthrough of deep convolutional architectures in computer vision catalyzed a new phase in remote sensing interpretation. CNNs introduced automatic feature learning, hierarchical abstraction, and spatial invariance, fundamentally transforming feature extraction. The introduction of Fully Convolutional Networks (FCN) by Long, Shelhamer, and Darrell (2015) established the foundation for dense prediction tasks such as semantic segmentation. Ronneberger, Fischer, and Brox (2015) further extended this framework through U-Net, a symmetrical encoder–decoder network designed to capture both global context and fine boundary details. Originally developed for biomedical image segmentation, U-Net’s skip connections enabled precise reconstruction of thin structures, making it particularly suitable for linear features like roads. Subsequent adaptations—DeepLab v3+ (Chen et al., 2018), PSPNet (Zhao et al., 2017), and D-LinkNet (Zhou et al., 2018)—incorporated multiscale context aggregation, dilated convolutions, and spatial pyramid pooling to enhance feature representativeness.

These advancements quickly transferred to remote sensing applications. Liu, Song, and Quan (2017) applied CNNs to classify road pixels from high-resolution aerial images, achieving substantial improvement over SVM and morphological methods. Lin et al. (2021) integrated atrous convolutions into U-Net to expand receptive fields without increasing parameters, successfully capturing both narrow and wide roads in complex scenes. Zhao et al. (2024) introduced FECF-Net, embedding feature-consistency modules to improve the uniformity of road segments across varying brightness levels.

2.2.2 Multi-scale and attention-enhanced architectures

As CNN-based models matured, researchers recognized that a single convolutional receptive field could not effectively handle roads of varying widths or complex intersections. Multi-scale feature fusion thus became a primary focus. Dai, Chang, and Li (2025) proposed the Dense Multi-scale U-Net (DMU-Net), which stacked multi-scale convolutional blocks to enhance information flow and reduce missing detections in mountainous areas. Similarly, Shao, Qi, and Zhang (2025) developed DESSNet, combining a Boundary Enhancement Residual Convolution Layer (BERCL) with a Shallow Content-Guided Spatial Attention Module (SCGSA) and a Multi-scale Context Selection Module (MSCSM). Their architecture significantly improved continuity and robustness under challenging illumination conditions. Attention mechanisms, first popularized by Vaswani et al. (2017) in the Transformer architecture, have also been adopted in remote sensing segmentation. Channel and spatial attention modules allow networks to focus selectively on road-related features while suppressing irrelevant background noise. Hybrid CNN–Transformer models have demonstrated superior long-range dependency modeling, further enhancing the recognition of thin and extended road segments.

2.2.3 Comparative studies on optimization strategies

While network architectures have received extensive attention, comparatively fewer studies have examined the influence of training optimization components—such as loss functions and optimizers—on segmentation outcomes. Yet, these factors play a crucial role in achieving stable convergence and accurate boundary prediction.

The Binary Cross-Entropy (BCE) loss function remains the most widely used for binary segmentation tasks because it directly minimizes the Kullback–Leibler divergence between predicted and target probability distributions. However, BCE treats all pixels equally, which can bias training toward majority classes when class imbalance exists. To address this, Dice Loss and its variants (e.g., Focal Loss, Tversky Loss) have been introduced to emphasize overlapping regions and mitigate imbalance (Sudre et al., 2017). Combining BCE and Dice Loss has proven effective for road segmentation, yet the trade-off between convergence stability and sensitivity to small targets remains underexplored.

Optimizer selection also profoundly affects model performance. Stochastic Gradient Descent (SGD) with momentum provides stable updates but requires careful learning-rate scheduling. Adaptive methods such as Adam (Kingma & Ba, 2015), AdamW (Loshchilov & Hutter, 2019), and RMSprop (Tieleman & Hinton, 2012) dynamically adjust learning rates based on historical gradient statistics, accelerating convergence on non-stationary objectives. Despite their widespread use, comprehensive comparative analyses under identical remote sensing datasets are rare. The present study addresses this gap by systematically quantifying the effects of both losses and optimizers within a controlled experimental design.

3. Methodology

The methodological framework of this study was designed to systematically evaluate deep-learning-based road extraction under different imaging conditions, spatial resolutions, and optimization strategies. It integrates a dual-source dataset, a convolutional neural network (CNN) architecture, a comparative loss-function–optimizer analysis, and comprehensive performance evaluation metrics. The overall workflow follows four main stages: data acquisition and preparation, model construction, training configuration, and accuracy assessment.

3.1 Study area and data sources

A. Study area description

The research area covers Zhanggong District, located in the city of Ganzhou, Jiangxi Province, China. This region lies between $25^{\circ}40'16''$ – $25^{\circ}58'56''$ N and $114^{\circ}46'44''$ – $115^{\circ}03'40''$ E, occupying approximately 375 km². The landscape is dominated by low hills and valleys, intersected by dense transportation networks that include national highways, rural roads, and river-crossing bridges. The subtropical monsoon climate ensures frequent vegetation cover and strong illumination variability, posing a challenging test bed for road extraction algorithms.

B. Data sources

Two categories of high-resolution imagery were used. Satellite imagery with 0.5 m spatial resolution was obtained from the Third National Land Survey dataset. UAV imagery with 0.05 m resolution was captured using a quad-rotor platform equipped with a high-definition RGB sensor.

These two data sources complement each other: the satellite imagery provides broad spatial coverage, while the UAV imagery offers detailed local textures suitable for fine-grained analysis.

3.2 Data Pre-processing and annotation

A. Radiometric and geometric corrections

All images underwent radiometric normalization to unify brightness and contrast, followed by orthorectification using ground-control points to ensure sub-pixel alignment between optical scenes. Histogram equalization was applied to enhance local contrast and facilitate edge detection.

B. Noise suppression

Median and guided filtering were employed to remove high-frequency noise, preserving sharp edges while reducing speckle effects. This step was crucial for UAV data, where sensor vibration occasionally introduced micro-blurring.

C. Manual annotation

Because public datasets covering the study area were unavailable, roads were manually vectorized using ArcGIS Pro. Annotators delineated road centerlines and edges with consistent width standards to produce accurate binary masks. The resulting shapefiles were rasterized at native resolutions.

D. Patch generation and dataset partitioning

The imagery and corresponding masks were cropped into 512×512 px tiles using the FeatureStation AI toolkit. After quality inspection, 6 813 patches were retained: 4 611 from satellite imagery and 2 202 from UAV imagery. The dataset was split into training (70 %), validation (20 %), and testing (10 %) subsets to balance learning and evaluation requirements.

E. Data augmentation

To enhance generalization and mitigate over-fitting, each training tile was subjected to random geometric and photometric transformations, including rotations ($0-90^\circ$), horizontal/vertical flips, brightness adjustment ($\pm 15\%$), Gaussian-noise injection, and scaling between 0.8-1.2. These augmentations simulate diverse viewing conditions and improve model robustness to illumination and orientation changes.

3.3 Model architecture

A. Network overview

The proposed model adopts a modified U-Net structure, consisting of an encoder–decoder framework with skip connections that merge low-level spatial information and high-level semantic features. This architecture was selected because of its proven ability to preserve boundary details and recover thin linear structures.

B. Encoder design

The encoder comprises four convolutional blocks, each containing two 3×3 convolutional layers followed by a ReLU activation and a 2×2 max-pooling layer. The number of feature maps doubles after each block ($64 \rightarrow 128 \rightarrow 256 \rightarrow 512$), enabling progressive abstraction of spatial features.

C. Bottleneck and feature fusion

At the deepest level, a bottleneck layer with 1 024 filters captures complex contextual information. Dilated convolutions are introduced to enlarge receptive fields without sacrificing

resolution. Residual connections are applied to stabilize gradient propagation.

D. Decoder and up-sampling

The decoder mirrors the encoder, employing transposed convolutions for up-sampling. Skip connections concatenate encoder features with corresponding decoder layers, ensuring spatial precision. Each up-sampling step halves the number of feature channels.

A. Output layer

The final layer uses a 1×1 convolution followed by a Sigmoid activation to produce a probability map representing the likelihood of each pixel belonging to the road class. Thresholding at 0.5 converts probabilities into binary predictions.

F. Model efficiency considerations

Batch normalization and dropout (rate = 0.2) are implemented to accelerate convergence and prevent over-fitting. Compared with heavier architectures such as DeepLab v3+ or PSPNet, this lightweight CNN offers a favorable trade-off between accuracy and computational cost, enabling near-real-time inference on a single GPU.

3.4 Loss function formulation

A. Binary Cross-Entropy (BCE) Loss

BCE measures the divergence between predicted probabilities and ground-truth labels. For each pixel i ,

$$L_{\text{BCE}} = -[y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$$

This function treats all pixels equally, providing stable gradients during early training stages.

B. Dice Loss Dice Loss emphasizes region overlap and alleviates class imbalance by maximizing the intersection between prediction and ground truth:

$$L_{\text{Dice}} = 1 - \frac{2 \sum_i p_i y_i + \epsilon}{\sum_i p_i + \sum_i y_i + \epsilon}$$

Here ϵ prevents division by zero. Dice Loss is sensitive to boundary accuracy but may exhibit unstable gradients when road pixels occupy only a small portion of the image.

C. Combined Loss Function To leverage complementary strengths, BCE and Dice Loss were combined as

$$L = \alpha L_{\text{BCE}} + \beta L_{\text{Dice}}$$

where $\alpha = \beta = 0.5$ after empirical tuning. This hybrid formulation balances pixel-wise accuracy with regional coherence.

3.5 Optimization algorithms

A. Stochastic Gradient Descent (SGD)

SGD updates model parameters by moving along the negative gradient of the loss function. Although simple and memory-efficient, it requires careful learning-rate scheduling and exhibits slow convergence on complex non-convex surfaces.

B. Adaptive Moment Estimation (Adam)

Adam (Kingma & Ba, 2015) combines momentum and adaptive learning-rate mechanisms, maintaining exponentially decaying averages of past gradients and their squares. It accelerates convergence but can overfit small datasets if learning rates decay too rapidly.

C. Adam with Weight Decay (AdamW)

AdamW (Loshchilov & Hutter, 2019) decouples weight decay from gradient updates, improving generalization and preventing over-parameterization. This optimizer has become standard for Transformer-based architectures and is tested here for comparison.

D. Root Mean Square Propagation (RMSprop)

RMSprop (Tieleman & Hinton, 2012) adjusts learning rates by dividing the gradient by an exponentially weighted moving average of its magnitude. It performs well in non-stationary objectives such as remote-sensing segmentation, offering smooth and stable convergence.

E. Optimizer configuration

All optimizers were initialized with a learning rate of 2×10^{-3} . A cosine-annealing scheduler gradually reduced the learning rate to 1×10^{-5} . Early stopping (patience = 10 epochs) halted training when validation loss ceased improving.

3.6 Training setup

Training was conducted on a workstation equipped with an Intel Xeon Silver 4214R CPU, 256 GB RAM, and an NVIDIA RTX 3090 GPU (24 GB VRAM). The implementation used Python 3.10 and PyTorch 2.1 within the FeatureStation AI ecosystem developed by the Chinese Academy of Surveying and Mapping. The batch size was set to 16; the number of epochs to 500. Weight initialization used the He-normal distribution. All experiments were repeated three times to ensure consistency, and mean results were reported. Training and validation curves were monitored to assess convergence behavior under different optimizer–loss combinations. After training, the model with the lowest validation loss was applied to the test set for final evaluation.

3.7 Performance evaluation metrics

Pixel Accuracy (PA) represents the ratio of correctly classified pixels to the total number of pixels:

$$PA = \frac{TP + TN}{TP + FP + TN + FN}$$

Intersection over Union (IoU) measures the overlap between predicted and reference masks:

$$IoU = \frac{TP}{TP + FP + FN}$$

Precision and Recall: Precision quantifies the proportion of correctly predicted road pixels among all predicted roads, while Recall assesses the proportion of true road pixels correctly identified:

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN}$$

The harmonic mean of Precision and Recall provides a balanced evaluation of both omission and commission errors:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

To ensure reliability, all results were averaged over the test tiles, and standard deviations were computed. Visual comparisons were supplemented with quantitative indicators to evaluate boundary fidelity, connectivity, and robustness under varying illumination conditions.

4. Results and discussion

4.1 Overview of experimental design

To ensure a fair comparison among training configurations, all experiments were conducted under identical hardware, software, and hyperparameter settings described previously. Each training run lasted 500 epochs with a batch size of 16. Early stopping based on validation loss prevented over-fitting. For every combination of loss function and optimizer, quantitative metrics—including Pixel Accuracy (PA), Intersection over Union (IoU), Precision, Recall, and F1-score—were computed on both satellite and UAV test sets. The statistical averages reported below represent the mean of three independent trials.

Overall, the CNN model demonstrated rapid convergence within 60–80 epochs across all configurations, with stable learning curves and minimal over-fitting. Visual inspection of predicted masks confirmed that the network successfully delineated most road surfaces, even under challenging conditions such as shadow occlusion and vegetation overlap. However, differences among loss functions and optimizers were evident in both convergence stability and final accuracy, as detailed in the following subsections.

4.2 Effect of loss function on model convergence

Two loss configurations—Binary Cross-Entropy (BCE) and BCE + Dice—were compared to examine their influence on learning behavior. Figure 4-1 illustrates the typical training and validation loss trajectories for each function (figure omitted in this text version).

During the early training phase, the BCE loss exhibited smoother and faster convergence. Validation loss decreased steadily, reaching a stable plateau after approximately 50 epochs. In contrast, the combined BCE + Dice loss presented oscillations during the first 30 epochs, indicating greater sensitivity to class imbalance and local gradients. Despite its theoretical advantage in emphasizing overlapping regions, the Dice component sometimes over-compensated for minor misalignments, producing unstable updates when road pixels accounted for less than 10 % of the total image area.

Quantitatively, BCE achieved slightly higher accuracy across most metrics. On the satellite dataset, BCE yielded an IoU of 80.8 % and an F1-score of 89.38 %, compared with 79.2 % and 88.17 % for the combined loss. The difference was even more pronounced on UAV imagery, where BCE produced clearer edge definitions and fewer discontinuities. These findings suggest that, for high-resolution road extraction tasks dominated by thin linear features, a pixel-wise entropy-based criterion provides sufficient gradient information without the instability sometimes induced by overlap-based losses.

Nevertheless, the Dice component remains valuable in other contexts involving large target areas or severe class imbalance. Future hybrid formulations might dynamically adjust the weighting between BCE and Dice according to the ratio of road to background pixels in each batch.

4.3 Comparative performance of optimizers

The choice of optimizer strongly influenced convergence dynamics and segmentation quality. Table 4-1 summarizes the quantitative performance achieved by four optimizers—SGD, Adam, AdamW, and RMSprop—using the BCE loss function.

Table 4-1: Quantitative comparison of segmentation performance using different optimization algorithms under identical training settings

Optimizer	PA (%)	IoU (%)	F1-score (%)	Precision (%)	Recall (%)
SGD	96.78	57.26	72.82	93.64	59.57
Adam	98.41	80.11	88.95	89.77	88.16
AdamW	98.19	77.76	87.49	87.51	87.46
RMSprop	98.49	80.80	89.38	91.30	87.54

The RMSprop optimizer achieved the best overall performance across all metrics. Its adaptive learning-rate adjustment effectively balanced convergence speed and stability, particularly in non-stationary loss landscapes characteristic of complex urban imagery. While Adam converged slightly faster during early epochs, it occasionally produced oversmoothed masks around narrow roads, possibly due to aggressive parameter updates. AdamW offered better generalization than Adam on the validation set but underperformed slightly in recall. SGD, although achieving high precision, suffered from low recall because of its limited capacity to escape sharp local minima.

Visual comparison corroborated these quantitative results. RMSprop-trained models produced cleaner and more continuous road segments with minimal false positives along rooftops or riverbanks. Consequently, subsequent experiments and discussions focus on the BCE + RMSprop configuration as the optimal baseline.

4.4 Influence of spatial resolution

To assess the impact of data resolution on model performance, separate models were trained on satellite (0.5 m) and UAV (0.05 m) imagery using the same network architecture and optimizer. The comparative results are shown in Table 4-2.

Table 4-2: Performance comparison between satellite and UAV imagery using the BCE + RMSprop configuration

Data source	PA (%)	IoU (%)	F1-score (%)	Precision (%)	Recall (%)
Satellite imagery	98.49	80.80	89.38	91.30	87.54
UAV imagery	96.43	84.53	91.62	92.45	90.80

The results reveal complementary strengths between the two data sources. Satellite imagery, despite lower spatial resolution, achieved the highest pixel accuracy because of its spectral consistency and wider spatial coverage. UAV imagery, on the other hand, excelled in IoU, F1-score, and recall, confirming that higher spatial detail enhances the model’s ability to delineate fine boundaries and small local roads.

However, UAV images exhibited slightly lower overall pixel accuracy due to noise, illumination variation, and surface reflections. This implies that higher resolution does not automatically guarantee better global accuracy; instead, its benefit lies primarily in improving connectivity and edge precision. The dual-source comparison further suggests that integrating multi-resolution data could leverage the stability of satellite imagery and the spatial detail of

UAV scenes for optimal performance.

4.5 Scene-specific performance evaluation

To understand contextual behavior, test samples were categorized into four scene types: (1) dense urban, (2) suburban residential, (3) vegetation-dominated, and (4) mountainous terrain. Quantitative evaluation showed that IoU values varied from 78 % in urban areas to 86 % in suburban scenes, reflecting differences in texture complexity and occlusion levels. The vegetation-dominated class exhibited the highest recall (91 %) but slightly lower precision (88 %), suggesting occasional confusion between roads and bare soil. Mountainous regions produced the lowest IoU (74 %) due to shadow-induced fragmentation.

These results emphasize that contextual adaptation remains essential. Incorporating digital elevation models (DEMs) or topographic priors could mitigate errors in steep terrain by constraining road continuity along feasible slopes.

4.6 Comparison with existing methods

To benchmark the proposed framework, results were compared with three representative baselines from the literature: (1) a traditional morphological filter + edge-linking method, (2) an SVM classifier using texture features, and (3) a standard DeepLab v3+ model trained under identical data conditions. Quantitative comparison is summarized in Table 4-3.

Table 4-3: Benchmark comparison between the proposed CNN model and representative baseline methods for high-resolution road extraction

Method	IoU (%)	F1-score (%)	Precision (%)	Recall (%)
Morphological + Edge-linking	58.3	73.6	81.2	67.3
SVM (texture-based)	65.8	79.4	84.0	75.2
DeepLab v3+	78.9	87.8	88.6	87.1
Proposed CNN (BCE + RMSprop)	84.5	91.6	92.4	90.8

The proposed approach surpasses all baselines by a significant margin, achieving up to 6–7 percentage-point improvement in IoU over DeepLab v3+. These gains stem from the model’s enhanced balance between spatial detail preservation and stability during training. The relatively small network size also ensures lower computational cost compared with heavy encoder–decoder designs, making it suitable for near-real-time mapping applications.

5. Conclusion and recommendations

This research presented a comprehensive investigation of deep-learning-based road extraction from high-resolution remote-sensing imagery, emphasizing the interplay among loss functions, optimization strategies, and data resolution. By constructing a unified experimental framework encompassing dual-source datasets—satellite (0.5 m) and UAV (0.05 m)—and employing a U-Net-derived CNN, the study systematically quantified how algorithmic choices influence segmentation performance.

The principal conclusions are summarized below:

A. CNN-based segmentation proved highly effective for delineating complex road networks in heterogeneous environments. The proposed model achieved an IoU of 80.8 % and F1 of 89.38 % on satellite imagery, and IoU of 84.53 % with F1 of 91.62 % on UAV imagery. These results substantially exceed those of classical machine-learning and morphological baselines, confirming that end-to-end deep feature learning overcomes the dependence on handcrafted descriptors.

B. Binary Cross-Entropy (BCE) produced more stable and rapid convergence than the combined BCE + Dice formulation. While Dice Loss can mitigate class imbalance, it exhibited gradient oscillation when the target proportion was small. Consequently, a pixel-wise entropy loss was sufficient to capture road boundaries accurately under high-resolution conditions.

C. Among the four tested optimizers, RMSprop delivered the most balanced performance in accuracy, recall, and convergence stability. Its adaptive learning-rate adjustment effectively handled the non-stationary optimization landscape of HRRSI segmentation, outperforming Adam, AdamW, and SGD.

D. Spatial resolution markedly affects segmentation outcomes. UAV imagery enhanced boundary delineation and connectivity but was more sensitive to local noise. Satellite imagery offered better global consistency and higher pixel accuracy. Their complementary characteristics suggest that fusing multi-resolution sources may yield optimal accuracy.

E. Errors mainly originated from annotation ambiguity, geometric mis-registration, and limited receptive-field size. CNNs capture local context effectively but struggle with long-range topology, occasionally producing disconnected road segments.

This study provides an empirically validated and theoretically grounded framework for CNN-based road extraction from high-resolution remote-sensing imagery. The findings confirm that: Deep learning substantially enhances extraction precision and robustness; BCE loss with RMSprop optimization offers an optimal balance between accuracy and stability; Spatial resolution significantly shapes performance characteristics; Further integration of attention mechanisms and topology-aware refinement will push performance toward near-perfect continuity.

The proposed methodology contributes both methodological clarity and practical guidelines for future geospatial artificial-intelligence research. As remote-sensing sensors continue to advance in resolution and coverage, and as deep networks evolve toward lighter and more interpretable forms, automated road extraction will play an increasingly central role in smart-city planning, autonomous-vehicle navigation, and disaster-response systems.

References

- Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4), 834-848.
- Dai, J., Chang, J., & Li, W. (2025). A dense multi-scale U-Net for mountainous road extraction from high-resolution remote sensing images. *Journal of Mapping Science*, 50(4), 92-102.
- Das, S., Mirnalinee, T. T., & Varghese, K. (2011). Use of salient features for the design of a multistage framework to extract roads from high-resolution multispectral satellite images. *IEEE Transactions on Geoscience and Remote Sensing*, 49(10), 3906-3931.
- Gaetano, R., Zerubia, J., & Scarpa, G. (2011). Morphological road segmentation in urban areas from high-resolution satellite images. *IEEE DSP Conference Proceedings*, 1-6.
- Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*.
- Lin, N., Zhang, X., & Wang, L. (2021). A road extraction method based on dilated U-Net for high-resolution remote sensing images. *Journal of Surveying and Mapping Science*, 46(9), 109-114.
- Liu, R., Song, J., & Quan, Y. (2017). Automatic road extraction from high-resolution remote sensing images using convolutional neural networks. *Journal of Xi'an University of Electronic Science and Technology*, 44(1), 100-105.
- Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. *International Conference on Learning Representations (ICLR)*.
- Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 234-241.
- Senthilnath, J., Rajeshwari, G., & Omkar, S. N. (2009). Automatic road extraction from high-resolution satellite images using texture progressive analysis. *ISPRS Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 38, 173-178.
- Shao, P., Qi, C., & Zhang, Y. (2025). A detail-enhanced and scale-selective network for road extraction from high-resolution remote sensing images. *Laser & Optoelectronics Progress*, 1-18.
- Shanmugam, L., & Kaliaperumal, V. (2016). Junction-aware water-flow approach for urban road network extraction. *IET Image Processing*, 10(3), 227-234.
- Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., & Cardoso, M. J. (2017). Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations. *Deep Learning in Medical Image Analysis (DLMIA)*, 240-248.
- Tieleman, T., & Hinton, G. (2012). Lecture 6.5 – RMSProp: Divide the gradient by a running average of its recent magnitude. COURSE: Neural Networks for Machine Learning.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I.

- (2017). Attention is all you need. *Advances in Neural Information Processing Systems (NIPS)*, 30, 5998-6008.
- Wang, W., Yang, N., & Zhang, Y. (2016). A review of road extraction from remote sensing images. *Journal of Traffic and Transportation Engineering*, 3(3), 271-282.
- Wu, B. (2005). Road extraction from remote sensing images based on improved Snake model with genetic optimization. *Wuhan University Journal of Natural Sciences*, 10(6), 1123-1128.
- Zhao, X., Luo, F., & Yang, H. (2024). Feature consistency perception network for road extraction from remote sensing imagery. *Laser & Optoelectronics Progress*, 61(18), 265-275.
- Zhou, L., Zhang, C., & Wu, M. (2018). D-LinkNet: LinkNet with pretrained encoder and dilated convolution for road extraction. *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 182-186.
- Zhu, E., Song, K., & Dai, J. (2016). SVM-MRF-based urban road extraction from high-resolution remote sensing images. *Remote Sensing Letters*, 7(9), 845-854.