JOURNAL OF BASIC SCIENCE AND ENGINEERING

Progressive Resolution Training with SWA and Test-Time Augmentation for Robust Road Segmentation in Remote Sensing Imagery

Nidhi Singh[1*], Aditi Sharma[2], Divyanshu Chauhan[3]

[1]Centre for Advanced Studies, Dr. A.P.J. Abdul Kalam Technical University, Lucknow, India
[2]Institute of Engineering and Technology, Lucknow, India
[3]Centre for Advanced Studies, Dr. A.P.J. Abdul Kalam Technical University, Lucknow, India.

DOI: https://doie.org/10.10399/JBSE.2026538738

ABSTRACT

Accurate extraction of road networks from high-resolution satellite imagery is inherently difficult for use in spatial applications due primarily to their complex topography and the extreme class imbalance associated with them, which typically leads to fragmentation of the resultant road segments. In order to overcome these structural challenges, we propose a U-Net multi-stage encoder-decoder architecture, which as part of a multi-encoder benchmark for evaluating the performance of several different architectures, uses the Focal Tversky Loss as a loss function to balance the extreme class imbalances associated with road networks and to provide increased penalties for false negatives. The proposed methodology employs a progressive resolution training approach, which has been established to improve generalization performance, as well as provide a significant increase in late-stage convergence when combined with Stochastic Weight Averaging (SWA), as well as to include SWA as part of the optimization process during the training phase. Finally, during the inference phase, all geometrically transformed predictions will be aggregated through a full Test-Time Augmentation (TTA), with an additional post-processing step, in which morphological operations will be applied to each prediction, in order to ensure the continuity of structures and remove noise artifacts from the predictions. The experimental results show that the proposed framework outperformed all other benchmark systems and confirmed through an ablation study that each incrementally added component contributed to the overall increase in performance, when compared to a baseline model. The most effective EfficientNet-B7 configuration results in an Intersection over Union (IoU) of 82.99%, a Dice coefficient of 90.68%, and F2 score 90.30%, thus providing an extremely accurate automated road network extraction approach with great geometric consistency.

1. INTRODUCTION

1.1 Background

Accurate extraction of roads from high-resolution remote sensing images can be extremely valuable for a variety of critical spatial applications such as urban planning, monitoring traffic flow, responding to disasters, or monitoring the environment. As a result of these continued breakthroughs in computational intelligence, the emergence of deep learning technology, like convolutional neural networks (CNNs), has greatly increased the ability and speed at which these networks can be extracted from satellite images and their accuracy and spatial precision when used to produce their outputs [1].

1.2 Shortcomings in Traditional Methods

While considerable progress has been made, automated road extraction from high-resolution satellite imagery remains fundamentally constrained by the numerous environmental topologies and visual occlusions of buildings, trees, and shadow artifacts that often lead to fragmented and discontinuous segmentations [2]. Standard deep learning architectures commonly experience semantic gaps and a loss of spatial information as the result of multiple down-sampling operations performed to achieve multiscale, long-range dependencies and fine boundary definition [3]. In addition, formulating the problem as a binary classification creates a significant issue with class imbalance, as relatively few pixels in comparison (thin roads) account for a very small portion of the total optical footprint [4]. As such, naive convolutional models are heavily biased to retrieve the background features, preserving no continuity in topology, and frequently producing inconsistent and noisy geometries across the highly heterogeneous urban landscape [5].

1.3 Motivation for Study

Advanced convolutional models are able to assess fine-grained features but the static scale networks struggle to learn long-range topological relationships and they also are unable to resolve extreme intra-class scale changes that result in prohibitively high computational costs [6]. Moreover, typical segmentation frameworks do not provide an adequate optimization objective for penalizing the occurrence of structural fragments when class imbalance is high, so structurally fragmented segments will not converge and exhibit instability at the late convergence stage [7]. This research presents an adaptive segmentation framework to achieve accurate, continuous extraction of roads from complex remote sensing images, avoiding structural fragmentation and limitations associated with multi-scale.

1.4 Aim and Objectives

Our goal in conducting this research study is to develop a reliable and accurate method for semantic segmentation that can be used to identify complex road networks in remote sensing images as accurately as possible while minimizing the effects of spatial occlusion and class imbalance. The means of achieving this goal will involve the implementation of various modified U-Net variants that employ progressive resolution scaling, differential fine-tuning, stochastic weight averaging (SWA), and morphological processing. In addition, this study will also generate multiple benchmarks using state-of-the-art encoder architectures (ResNet-152, MiT-B0, MobileOne-S3, SENet154, EfficientNet-B7, InceptionV4) to provide comparisons of their relative model performances, feature extraction capabilities, and accuracy in delineating complex road networks.

1.5 Contribution of this Work

This study develops a new robust multi-level segmentation framework based on an encoder-decoder architecture. It uses the Focal Tversky Loss objective to specifically address the problem of significant class imbalance; it penalises false negatives. The primary contribution of this research is to create a training paradigm that increasingly progresses from lower-resolution input for context information to higher-resolution input for improving the boundaries through their fine-tuning. This will enable progressively improving the balance between the elements/levels of spatial scale on which the features of interest are being adapted to the available computational resources. Stochastic Weight Averaging (SWA) will also be implemented into the optimisation workflow to promote additional generalisation ability and to enhance late-stage convergence stability. In addition, by performing extensive Test-Time Augmentation (TTA) in combination with morphological post-processing techniques, the merged output from the transformed predictions will provide a consistent geometric structure and degree of precision between all of the road networks; therefore, providing reliable predictions throughout complex regions with minimal fragmentation.

2. LITERATURE REVIEW

2.1 Related Works

Table 1. Comparative analysis of existing techniques and methodologies for Road Network Extraction from Satellite Images.

S.No.	Author	Technology Used	Dataset Used	Key Findings	Performance
1	Liu et al. (2023) [1]	RoadFormer (Swin Transformer + CNN)	DeepGlobe, Mass. Roads	Uses Swin Transformer backbone with separable convolution to capture long-range global context.	IoU: 74.3% (DeepGlobe)
2	Zao et al. (2024) [3]	TopoRoad (Topology-Guided Learning)	CityScale, SpaceNet	Decouples extraction into vertex/orientation prediction to learn vectorized representations directly.	APLS: 68.5% (CityScale)
3	Sultonov et al. (2022) [8]	Mixer U-Net (ConvMixer + U-Net)	UAV Imagery	Replaces standard conv layers with ConvMixer to capture spatial-channel dependencies efficiently	IoU: 78.4% (UAV Dataset)

4	Tan et al. (2020) [9]	VecRoad (Iterative Graph Exploration)	RoadTracer Dataset	Uses flexible step sizes and segmentation cues to iteratively explore and vectorize road graphs.	F1: 82.6% (RoadTracer DS)
5	Bastani et al. (2018) [10]	RoadTracer (Iterative Graph Construction)	40 US Cities	First major method to extract road networks as a graph directly rather than pixels-then-vector.	Error rate: ~5% (Junction recovery)
6	Sun et al. (2018) [11]	Stacked U-Net (Hybrid Loss)	DeepGlobe	Stacks multiple U-Nets and uses a hybrid loss function to iteratively refine road segmentation.	IoU: 63.5% (DeepGlobe)
7	Abdollahi et al. (2020) [12]	VNet (Vectorization Network)	DeepGlobe	Proposes a method to extract vector road maps directly using a deep learning framework.	F1: 82% (DeepGlobe)
8	Xu et al. (2018) [13]	GL-Dense-U-Net: DenseNet + Attention	Google Earth	DenseNet+U-Net with Global/Local attention to fix breaks and refine edges.	F1-Score: 0.8877
9	Xin et al. (2019) [14]	DenseUNet: Dense Connection Units	Google Earth, Conghua	Dense connections in U-Net maximize flow and reduce vanishing gradients.	F1-Score: 0.91
10	He et al. (2020) [15]	Sat2Graph: Graph-Tensor Encoding	City-Scale (US), SpaceNet	Direct graph tensor prediction; handles overpasses better than pixel methods.	TOPO: Surpassed RoadTracer

2.2 Identified Research Gaps

There have been advances in automated extraction of road features from images, but there remains a significant gap between the pixel-based segmentation of the image and graph-based approaches to highway mapping. Recent advances in automated road extraction using deep learning networks designed for highway mapping such as U-Net, attention-based models show the ability to extract information from images across different levels of scale [2][3][8]. However, since these models use pixel-based methods to extract information from images, they typically lack the ability to retain continuity between distant components and frequently generate fragmented networks due to occlusions or shadows in the images used. Conversely, iterative graph extraction implementations of VecRoad [9], RoadTracer [10], Sat2Graph [15], focusing on achieving geometric accuracy and maintaining topology but have varying degrees of inability to scale computationally and to process high-resolution images containing multiple levels of spatial variability at once. As such, there is a large research gap where currently there is no integrated, end-to-end differentiable model that simultaneously captures dense semantic feature representations and includes explicit topological information on a global scale. Developing a method to bridge this gap will also be critical to ensuring the development of new fusion approaches that produce accurate spatially based localizations and maintain structural connectivity between multiple components of the complex urban built environment.

3. MATERIALS AND METHODS

3.1 Dataset Description

The TGRS-Road dataset comprises 224 aerial images captured via high-resolution satellite imagery. Each image has a resolution of 1.2m and a corresponding binary ground truth mask. There are three separate, non-overlapping datasets; 160 Training Images, 20 Validation Images and 44 Test Images. The data comprises high-resolution imagery captured by the same satellite and at a minimum starting size of 600 x 600 pixels using standard 3-channel RGB colour space. A binary road mask has been created from these images by collapsing the original raw semantic annotations into a single-channel (grayscale) array of road pixels in the corresponding road topology of between 12 and 15 pixels in width each represented as floating point values of 1.0 and background pixels represented as 0.0 that will create definitive classification targets to determine the presence of roads in a given image.

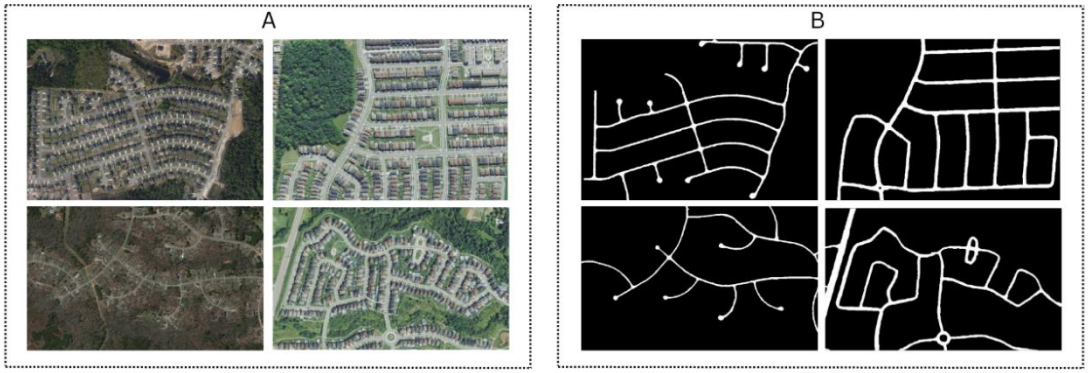

Fig. 1. High resolution satellite images and their corresponding mask (A) training image set (B) corresponding mask to training set

3.2 Robust Data Augmentation Pipeline

A complete set of augmentations was developed as part of the training pipeline for the model to ensure good feature extraction with minimal potential for overfitting by the model. The geometric augmentations applied to images in the training dataset were transformed both horizontally and vertically, particularly rotating an image by 90° (with a probability of 0.5) as shown in fig.2. It was expected that the model would learn how to recognize the same attributes from each version of the original image after it had seen the entire dataset consisting of geometric augmentations. Additional shifts/scales/rotates were applied to the training dataset with the following minimum/maximum percentages: shift = 6.25%, scale = 10%, rotate = 45%, p=50%. To add additional photometric variations, brightness/contrast (limit: 20%) and hue/saturation/value (hue: 10, saturation: 20, value: 10 and probability: 30%) were randomly altered. The method known as stochastic filtering was employed to implement two types of filters (i.e., Gaussian blur and sharpen) at random between two or more type of filters (Gaussian blur Kernel size = 3-5 and sharpen alpha = 0.2-0.5) which increased the characteristics and diversity in the training set due to differences in sensor output (0.2 probability of stochastic variance). A more detailed visual representation of the stochastic filtering can be found in fig. 3. Each hyperparameter had limits imposed to control potential distortions from the images during the augmentation process so as to encourage the model to learn invariant generalization of the topographic features irrespective of the degree of distortion experienced by the images.

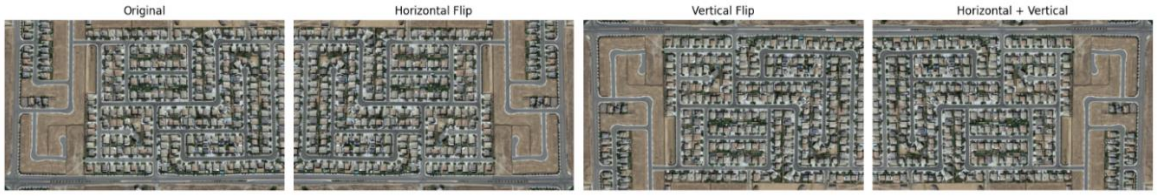

Fig. 2. Geometric Augmentations applied on Test Set

Fig. 3. Photometric Augmentations applied on Test Set

3.3 Overall Framework

To rigorously evaluate the framework, we have established training, validation and test data sets. The training and validation data have undergone successive scaling (512x512 pixels to 768x768 pixels) to speed up the time to initial convergence and, in later epochs, refine hyper-degree spatial detail. While performing U-Net optimizations, the training data went through extensive spatial and photometric augmentation; the validation data provides epoch-based generalization feedback. To provide additional robustness and stability, the final ten epochs of SWA were utilized (fig. 4.) during testing. Test results are based on an evaluation of the test data which will remain at a fixed

resolution of 768x768 pixels using the SWA-generated model as compared against the average model together with 8-fold TTA. Finally, morphological post-processing will correct topological anomalies and smooth the boundary edges to compute final evaluation metrics.

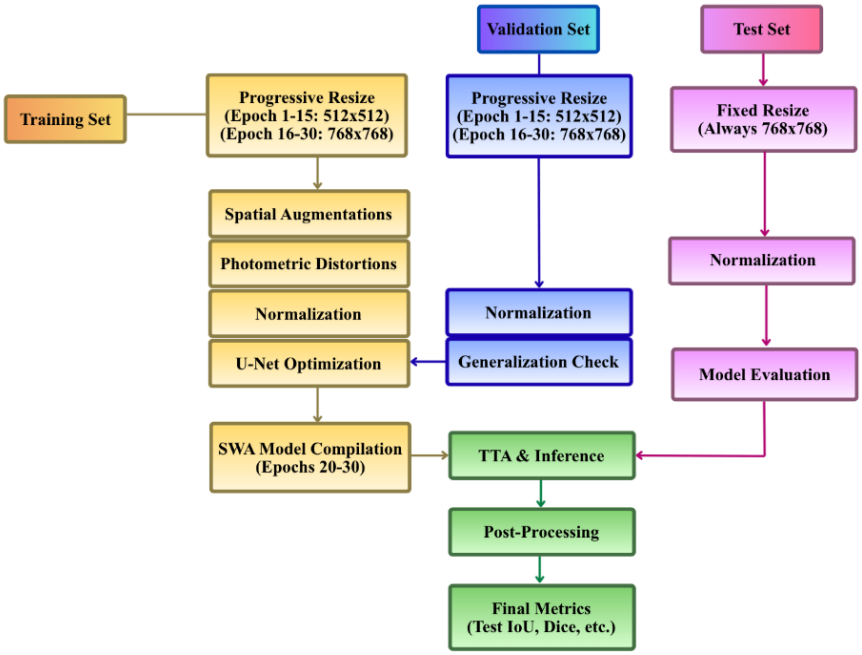

Fig. 4. Comprehensive flowchart of the end-to-end dataset processing and U-Net optimization pipeline, detailing progressive resizing strategies, Stochastic Weight Averaging (SWA), and Test-Time Augmentation (TTA).

3.4 Network Architecture

To achieve high-fidelity semantic segmentation of complex road networks, this study is focused on developing an accurate semantic segmentation system for complex road networks using the U-Net architecture. The U-Net consists of a symmetric encoder-decoder structure with lateral skip connections between all layers within the architecture. The lateral skip connections enable accurate reconstruction of spatial information because the decoder leverages higher-level semantic representation of objects from the point-of-view of an encoder to create lower-resolution spatial representation for reconstructing a physical environment. Therefore, through the use of lateral skip connections, U-Net can retrieve very fine details from a physical environment which can result from multiple applications of low-resolution data to produce individually small linear objects like narrow connected objects or sharply defined edges in the removal of remote sensing imagery.

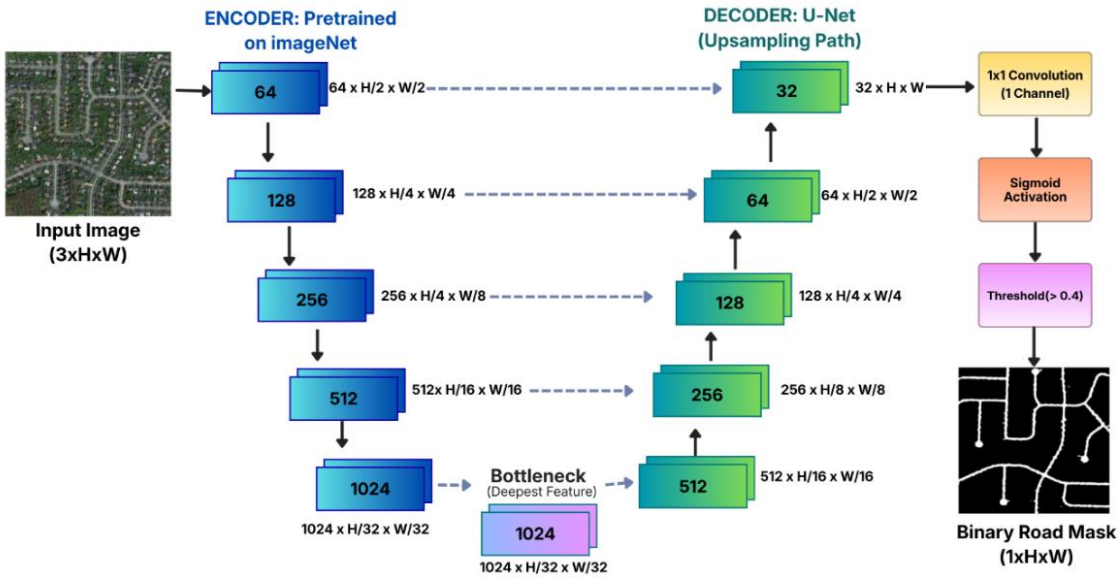

Fig. 5. U-Net segmentation architecture framework

Vol. 23, No. 1, (2026)
ISSN: 1005-0930

3.5 Encoder Architectures Evaluated

In this analysis, we observed how various backbone architecture and their operational functions perform. ResNet-152 is a deep residual learning-based backbone architecture that uses bottleneck building blocks to create identity shortcuts for communication of gradients between layers. MiT-B0 (Vision Transformer) uses self-attention to perform hierarchical attention on images by grouping patches into overlapping sets and then encoding positional locations for each patch rather than using a traditional encoding method. MobileOne-S3 uses structural re-parametrization of convolutional layers by training multiple parallel branches that mathematically collapse into the same branch at inference. SENet154 performs explicit modeling of channel dependencies between channels using squeeze and excite modules, and dynamically recalibrating the channel responses to extracted features. EfficientNet-B7 achieves scale by scaling network width, depth, input resolution equally for obtaining the best achievable network accuracy. Finally, InceptionV4 uses very highly parallel multi-branch modules with spatial factorization to allow processing images of different scales at the same time.

3.6 Progressive Resolution Training Strategy

The proposed progressive training method of segmentation images for future learning is designed to improve road segmentation performance. Phase 1 of training uses low-resolution (512 x 512) images to build the model by providing a basic understanding of the general spatial relationship between the targets and the approximate shape of the road. Phase 2 fine-tunes the model using higher-resolution (768 x 768) images to enable the model to extract the detailed boundaries of the segments it has segmented. The coarse-to-fine model allows for multi-resolution learning of similar features, while also providing the model with the ability to generalize across varying exposures and various levels of complexity corresponding to the width of the road itself. Thus, the two phases of training establish very stable transition points (optimization) from one phase to the second. Due to the transition between Phase 1 and Phase 2 optimization stables and spatial sensitivity increases.

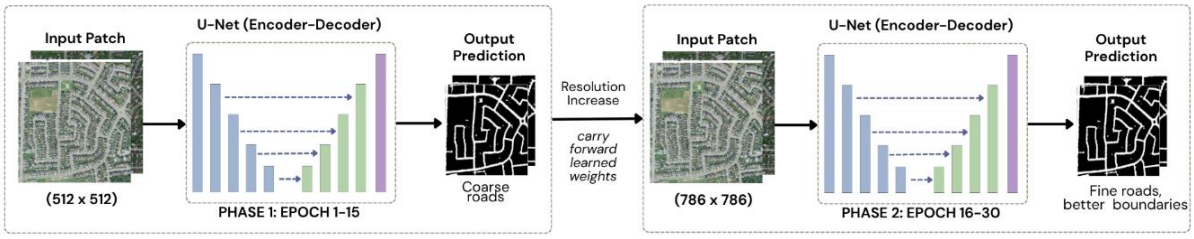

Fig. 6. Proposed two-phase training strategy. The model is first trained on lower-resolution images (512×512) to learn global road structures, followed by higher-resolution training (768×768) for fine-grained refinement.

3.7 Transfer Learning and Fine Tuning

In order to improve the precision of the system by facilitating the extraction of low-level features, all of the encoder layers were initialized with the weights of the encoders from pre-trained ImageNet dataset. In addition, the optimizer used for the decoder and segmentation layers was an AdamW optimizer combined with a differentially implemented learning rate schedule that decreases the chance of losing previously learned representations associated with pre-trained encoder weights during training as the AdamW uses an exponentially decreasing learning rate. As all encoder weights from pre-trained networks were fine-tuned with a very slow learning rate of $3x10^{-5}$, we could preserve all previously learned representations to include existing edges and textures but still be able to modify their functionality without overwriting them unnecessarily. Also, since the decoder and segmentation layer parameters were initialized randomly with a full-base learning rate of $3x10^{-4}$ prior to training; therefore, they could learn quickly with a considerably lower complexity and develop very complex relationships between the extracted features and ideal spatial road masks. Finally, all of the parameters were subject to a uniform weight decay regularization of 10^{-4} during training to penalize excess weight accumulation so as to reduce the likelihood of overfitting to the given data.

3.8 Loss Function Design

Our approach to the class imbalance issue that occurs when working with models, is to optimize using the Focal Tversky Loss ($\mathcal{L}_{\mathcal{FTL}}$).. Where N is the number of total elements (pixels) in an image and $p_i \in [0,1]$is the probability predicted by the i-th element of the image belonging to the target class given by the sigmoid activation function. Let $g_i \in \{0,1\}$ represent the corresponding binary ground truth label.

Let's now define the continuous approximations of confusion matrix elements (i.e., True Positives (TP), False Positives (FP), and False Negatives (FN)) as follows:

$$TP = \sum_{i=1}^{N} p_i g_i$$

$$FP = \sum_{i=1}^{N} p_i (1 - g_i)$$

$$FN = \sum_{i=1}^{N} (1 - p_i) g_i$$

The Tversky Index (T) is defined to measure the overlap between the prediction and the ground truth. It allows for unequal penalties for false positives and false negatives by allowing for different weighting of those components as follows:

$$T = \frac{TP + \epsilon}{TP + \alpha FP + \beta FN + \epsilon}$$

where α and β are tunable hyperparameters that control the penalty assigned to false positives and false negatives, respectively. A small stabilizing constant, ϵ (set to 10^{-7}), is added to both the numerator and denominator to prevent division by zero.

To further encourage the model to focus on hard, misclassified examples rather than easily segmented background regions, a focal parameter γ is applied to the complement of the Tversky Index. The final Focal Tversky Loss is formulated as:

$$\mathcal{L}_{\mathcal{FTL}} = (1 - T)^{\gamma}$$

By substituting the continuous component approximations into the equation, the complete objective function is expressed as:

$$\mathcal{L}_{\mathcal{FTL}} = \left(1 - \frac{\sum_{i=1}^{N} p_i g_i + \epsilon}{\sum_{i=1}^{N} p_i g_i + \alpha \sum_{i=1}^{N} p_i (1 - g_i) + \beta \sum_{i=1}^{N} (1 - p_i) g_i + \epsilon}\right)^{\gamma}$$

In our experiments, the hyperparameters are empirically set to $\alpha = 0.25$, $\beta = 0.75$, and $\gamma = 1.5$ to heavily penalize false negatives and prioritize learning on challenging class instances.

3.9 Optimization and Advanced Training Dynamics

The network optimization process utilized the AdamW optimizer as an optimization algorithm with a weight decay of 1x10^-4 considered in order to provide a strong level of regularization and penalize the model for having extremely large weights to help prevent model overfitting. To allow for gradual convergence, a learning rate scheduler that utilized cosine annealing to decay the learning rates over the entire training period T_{max}=30 was also incorporated into the training process. SWA (Stochastic Weight Averaging) was employed to enhance the generalization, on epoch 20 as the model had already converged at that point so that the parameters could be averaged over the final epochs while being trained with a slow learning rate (base LR/5) of approximately $6x10^{-5}$. The gradual nature of the weight decay provided a very stable local search space incorporating much broader, flatter, and more robust minimum loss functions for the network during the entire training process. Following the completion of the training process, the batch normalization was then applied on the training dataset in order to realign the internal statistics of the data being used with the weights of the completed SWA process.

AdamW Optimizer (Weight Decay Update),

$$\boldsymbol{\theta_t = \theta_{t-1} - \eta(\nabla L(\theta_{t-1}) + \lambda \theta_{t-1})}$$

Where $\boldsymbol{\theta}$ represents the network weights, $\boldsymbol{\eta}$ is the learning rate, $\mathbf{L}$ is the loss function, and $\boldsymbol{\lambda}$ is the weight decay coefficient.

Cosine Annealing Learning Rate Scheduler,

$$\eta_t = \eta_{min} + \frac{1}{2}(\eta_{max} - \eta_{min})\left(1 + \cos\left(\frac{T_{cur}}{T_{max}}\pi\right)\right)$$

Where η is the current learning rate, η_{max} and η_{min} are the boundary learning rates, T_{cur} is the current epoch, and T_{max} is the total number of epochs.

Stochastic Weight Averaging (SWA),

$$w_{SWA} = \frac{1}{n}\sum_{i=1}^{n} w_i$$

Where w_{SWA} represents the final averaged weights, n is the number of epochs in the SWA phase, and w_i are the weights at each specific epoch i.

3.10 Test-Time Augmentation (TTA)

During the inference stage, an 8-fold TTA was done by taking spatial flipping and 90-degree rotations of each data point so that each pixel would have stable probability map as determined by rotating the original location. Each of these rotated maps had a corresponding probability value assigned to its pixel and then all corresponding pixels from each rotated map were averaged to create a stable probability map that had reduced spatial bias and variance. A confidence threshold of 0.4 was selected to binarize the probability maps, resulting in the maximum recall possible for connections between faint and non-continuous roads prior to their being refined. Additionally, after using the thresholds to binarize the probability maps, advanced heuristics were applied to ensure the geospatial accuracy of the topographic structures that would ultimately reconstruct the roads (skimage & OpenCV). Objects with less than 200-pixel count were removed as isolated (False Positive) noise in the geospatial network. Furthermore, using a gap-filling methodology with threshold sizes of less than 200 pixels, False Negative discontinuities due to obstructions and shadows cast by trees were repaired. Finally, the road network was smoothed using a focus on morphology which resulted in an actual averaged variation of all connections created along the roads but limited the overall increase in the physical roadway widths.

3.11 Evaluation Metrics

Evaluation metrics at the pixel level were computed to provide a definitive evaluation of how well the new model predicts road surfaces. Evaluation metrics based on all elements in a confusion matrix: True Positive (TP); the number of road pixels that are classified accurately; False Positive (FP); the number of pixels classified as road, but are actually background; True Negative (TN); the number of background pixels classified accurately, and False Negative (FN); the number of road pixels the model failed to classify. An error constant $\epsilon = 1 \times 10^{-7}$ was added to all denominators as a safeguard against division by zero in the evaluation metrics calculations.

Intersection over Union (IoU), also known as the Jaccard Index, IoU is the primary benchmark for semantic segmentation. It measures the direct overlap between the predicted road mask and the ground truth, severely penalizing both false positives and false negatives,

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

The Dice Coefficient represents the harmonic mean of Precision and Recall. It is heavily utilized in medical and remote sensing domains because it provides a balanced evaluation of the model's performance, particularly when the region of interest (roads) occupies a small fraction of the total image area,

$$Dice = \frac{2TP}{2TP + FP + FN}$$

Vol. 23, No. 1, (2026)
ISSN: 1005-0930

Precision calculates the exactness of the model. It defines the proportion of pixels predicted as "road" that actually belong to the road class, highlighting the model's resistance to background noise.

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FP}}$$

Recall measures the completeness of the extraction. It quantifies the proportion of actual road pixels successfully identified by the model, which is critical for ensuring continuous network connectivity.

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}}$$

The F2-score is a weighted variation of the F-measure that places twice as much emphasis on Recall as it does on Precision. In the context of road extraction, higher recall is often prioritized over absolute precision, as it is generally easier to prune false positives during post-processing than to reconstruct missing, disconnected road segments,

$$\boldsymbol{F_2} = \frac{\mathbf{5} \times \boldsymbol{Precision} \times \boldsymbol{Recall}}{\mathbf{4} \times \boldsymbol{Precision} + \boldsymbol{Recall} + \boldsymbol{\epsilon}} = \frac{(\mathbf{1} + \mathbf{2^2}) \times \boldsymbol{TP}}{(\mathbf{1} + \mathbf{2^2}) \times \boldsymbol{TP} + \mathbf{2^2} \times \boldsymbol{FN} + \boldsymbol{FP} + \boldsymbol{\epsilon}}$$

Accuracy, this metric calculates the global ratio of correctly classified pixels (both road and background) to the total number of pixels. While provided for completeness, accuracy is often considered a secondary metric in road extraction due to the severe class imbalance, the vast majority of pixels are background $\boldsymbol{TN}$, which can artificially inflate the score,

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP} + \boldsymbol{TN}}{\boldsymbol{TP} + \boldsymbol{TN} + \boldsymbol{FP} + \boldsymbol{FN}}$$

4. EXPERIMENTAL SETUP

PyTorch with Segmentation Models (Pytorch) library has also been used as the basis for implementing the defined architecture, with all calculations being performed on GPU (CUDA) architectures. The training scenario with two different configurations requires dynamic batch size scaling as part of the methodology utilized to optimize the system for the training scenarios. A second benefit was to use stochastic data augmentation with the albumentations library to allow for the generalizability of the extracted features throughout the training process. Additionally, topological artifacts were removed based on probability relative to the inference and spatial masks were generated based on systematic methodologies using scikit image and opencv. Fixed seudo-random numbers generator seed was strictly maintained throughout experiment.

5. RESULTS AND DISCUSSION

5.1 Encoder Comparison Study

Table 2. Quantitative performance comparison of various encoder architectures.

Encoder	IoU (%)	Dice (%)	Precision (%)	Recall (%)	F2 (%)
Mobileone_s3	80.83 ± 3.35	89.97 ± 2.07	91.15 ± 6.69	89.55 ± 5.25	89.62 ± 3.15
Resnet152	81.86 ± 3.31	89.99 ± 2.05	91.76 ± 6.55	89.00 ± 5.39	89.30 ± 3.32
Densenet201	81.65 ± 3.02	89.87 ± 1.83	91.86 ± 6.01	88.63 ± 5.56	89.04 ± 3.54
Senet154	82.70 ± 3.00	90.50 ± 1.81	92.72 ± 5.41	89.01 ± 5.83	89.53 ± 3.92
Inceptionv4	82.09 ± 2.79	90.14 ± 1.68	92.11 ± 5.89	88.92 ± 5.60	89.32 ± 3.58
Mit_b0	82.79 ± 2.76	90.56 ± 1.66	91.68 ± 5.87	90.13 ± 5.52	90.22 ± 3.49
Efficientnet-b7	82.99 ± 3.03	90.68 ± 1.82	92.30 ± 6.17	89.79 ± 5.45	90.30 ± 5.10

5.2 Ablation Study

Table 3. Ablation study detailing the incremental impact of proposed methodology components on model performance.

Configuration	IoU (%)	Dice (%)	Precision (%)	Recall (%)
Baseline U-Net (resnet34)	75.96	86.26	94.67	79.93
U-Net + efficientnet-b7	75.68	86.07	95.79	78.78
+ TTA	82.18	90.19	91.11	90.00
+ SWA	82.39	90.31	91.83	89.56
+ Progressive Resizing	82.41	90.33	91.97	89.45
+ Post-processing	82.99	90.68	92.30	89.79

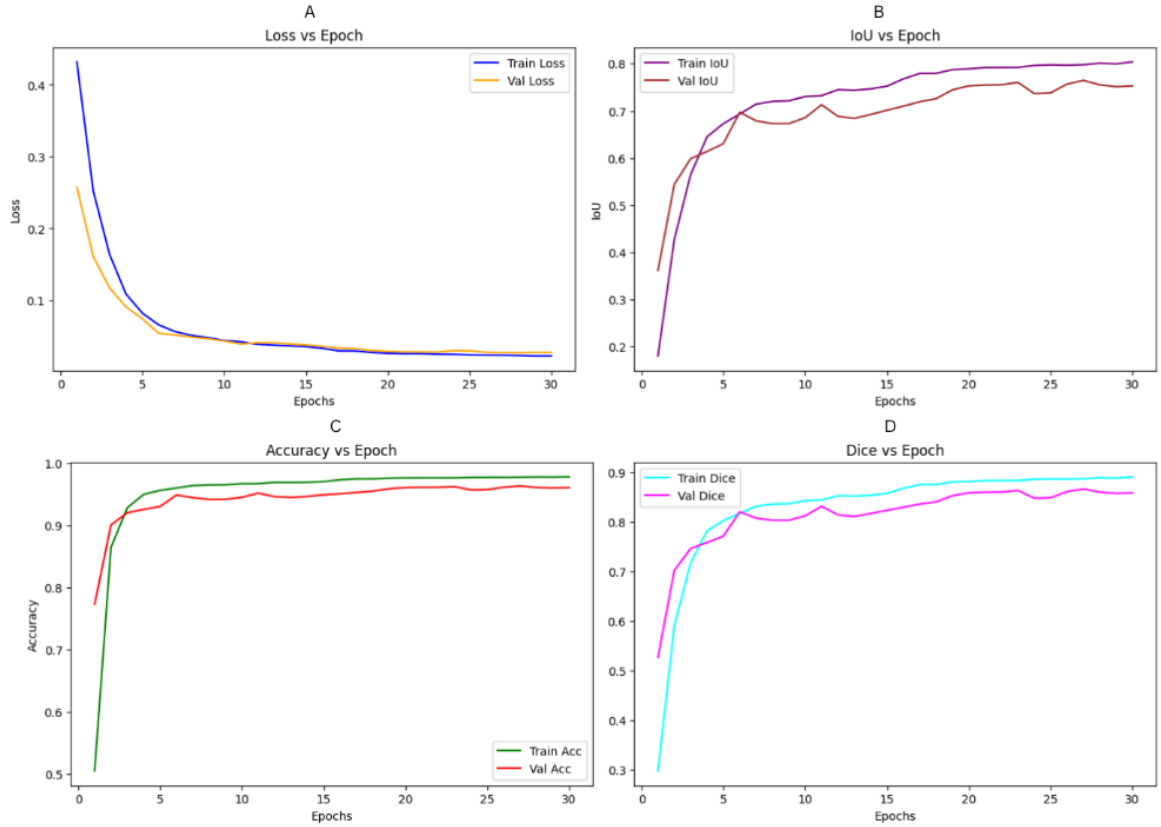

Fig. 7. Training and validation performance metrics across 30 epochs for the road network extraction model. The subplots illustrate the convergence and generalization trends evaluated through: (A) Focal Tversky Loss, (B) Intersection over Union (IoU), (C) Pixel-wise Accuracy, and (D) Dice Coefficient.

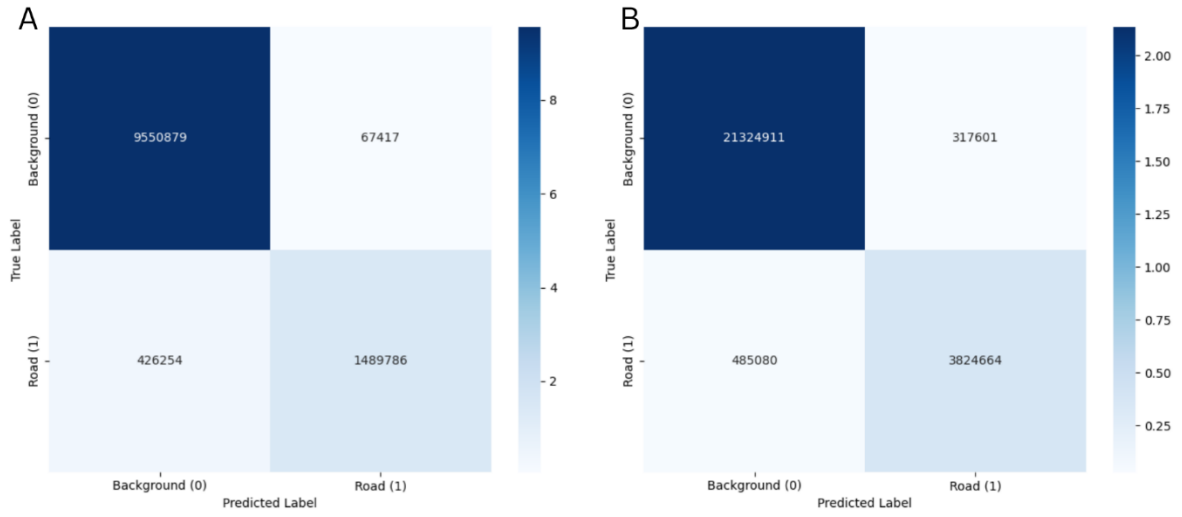

Fig. 8. Confusion matrix comparison of road segmentation performance: (A) baseline U-Net with ResNet34 encoder and (B) optimized EfficientNet-B7-based model with advanced enhancements

5.3 Comparison with Existing Methodologies

Table 4. Comparative analysis of the metrics IoU and Dice on the proposed framework against existing methodologies

Methodology / Architecture	IoU (%)	Dice (%)
Proposed efficientnet-b7	82.99	90.68
CoANet (2021) [17]	64.73	79.96
Joint Angle Prediction (2023) [18]	71.52	83.40
UBR-Net (2025) [2]	57.22	75.18
DeepLabV3+ Pipeline (2025) [3]	78.40	87.17
DSWFNet (2026) [19]	66.07	79.57
DS-Unet (2026) [7]	79.25	87.21

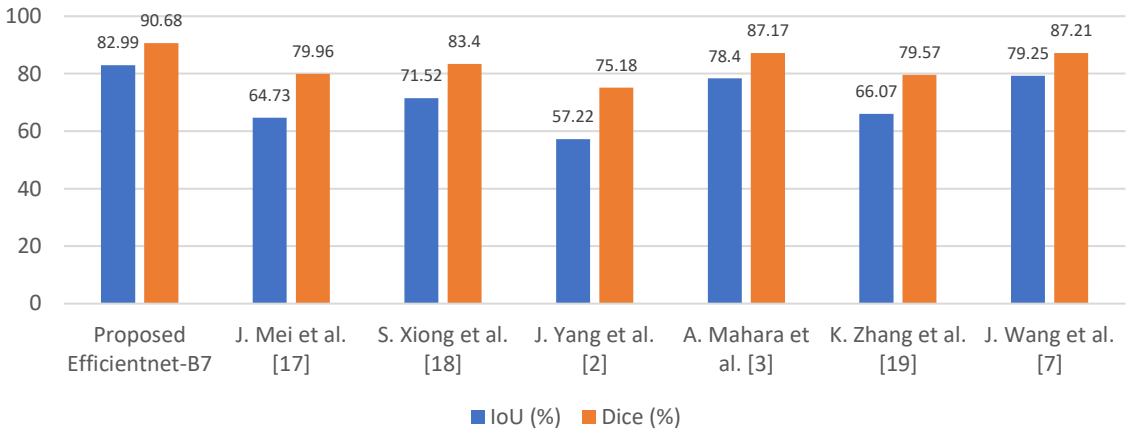

Fig. 9. Comparative analysis of the proposed framework against existing methodologies

5.4 Qualitative Result

A comparative visualization of four panels was used for qualitative evaluation. The layout compares the denormalized optical image to the ground truth mask and final refined prediction and annotates the final refined prediction with its respective Intersection over Union (IoU) score. To better understand area of classification failure, a spatial error map was created using a 'magma' colormap to illustrate the classification failure areas. This error map shows all False Positives (over-predicted areas) in high-intensity thermal values (yellow) and False Negatives (correct areas of under-prediction) in mid-intensity thermal values (purple), allowing for a detailed morphological analysis of the network's predictively poor performance.

Vol. 23, No. 1, (2026)
ISSN: 1005-0930

Fig. 10. Qualitative results of the proposed road extraction framework where (A) raw high-resolution resized input imagery (B) the corresponding ground truth annotations (C) the predicted binary segmentation mask (D) error map visualizing false positives (yellow) and false negatives (purple).

5.5 Impact of Optimization and Multi-Scale Learning

Using Stochastic Weight Averaging (SWA) to stabilize a late-stage convergence process ensures that the model being trained does not converge to a sharp-minima while still providing for improved generalization in novel urban topologies due to the wider variety of flatter local minima created by sharing weights over all previous training iterations. Furthermore, it has been demonstrated that stochastic weight averaging supplementary with progressive resolution training allows for progressive learning at greater distances from the view point, thereby giving the model a broader understanding of the relationship between global spatial contexts and the detailed boundaries of the segmented objects. By using this type of training method, the model will remain capable of maintaining long-range dependencies while correctly segmenting a fragmented roadway.

5.6 Inference Refinement and Robustness

Test-Time Augmentation (TTA) significantly reduces prediction variance by aggregating multiple geometric perspectives, effectively removes spurious noise artifacts. When coupled with morphological post-processing, the framework successfully mitigates structural fragmentation caused by visual occlusions, such as shadows and building overhangs. This is particularly vital for thin structures and areas where the model faces high annotation noise sensitivity, as the ensemble approach stabilizes the output against pixel-level irregularities.

6. LIMITATIONS

The framework proposed has a high Dice coefficient of 90.68% and an F2 score of 90.30%. However, there are still structural constraints to the model. The model shows local sensitivity to both annotation noise and very large or very small scales. Additionally, TTA and SWA introduce a lot of additional computation and cost time which results in longer inference times. Therefore, there is a trade-off between topological accuracy and real-time deployment in remote sensing applications especially in hardware constrained environments where low latency is essential.

7. CONCLUSION AND FUTURE WORK

In this research paper, a novel research methodology is being proposed for transforming road networks from HR satellite imagery utilizing morphological measurement metrics. The core contribution of the methodology used is to encompass a two-phase progressive resolution training strategy, the integration of Stochastic Weight Averaging (SWA) to stabilize late-stage convergence, and an 8-fold Test-Time Augmentation (TTA) pipeline. The Focal Tversky Loss function was employed in this research paper due to it addressing extreme imbalances in the number of classes and penalizing false negatives within remote sensed datasets. The high-performance model, which utilizes a U-Net backbone constructed with an EfficientNet-B7 backbone yielded the highest baseline results yielding both an Intersection Over Union (IoU) of 82.99% and a Dice Coefficient of 90.68% for the extraction of road networks. The created framework will be able to produce topologically accurate data, and thus, usable at a national Geographic Information System (GIS) level; and thereby potentially assisting in providing autonomous vehicles with accurate geo-spatial information for their navigation throughout the built environment.

In the future, there will be further investigations related to enhancing the performance of networks by improving their overall efficiency and structural integrity. More specifically, this project will create a means of enforcing strong connectivity of networks based on how well fragmented structures that have a graph-oriented topology maintain their integrity through the use of a preservation loss function; the results from this work will greatly affect this phase of the project. By using hybrid CNN-Transformer encoders, we will be able to both capture long range global dependencies and localized points of interest; both of these are possible to achieve using these two different architectural types. During this second phase of research, using semi-supervised and self-supervised methods of training with a large amount of unlabeled satellite imagery will also allow for a more efficient use of these types of datasets. Ultimately, if effective techniques for domain adaptation can be developed, then it will be possible to generalize this research across a variety of different types of satellite sensors and also across millions of targeted geographic areas.

REFERENCES

[1] R. Liu, J. Wu, W. Lu, Q. Miao, H. Zhang, X. Liu, Z. Lu, and L. Li, "A Review of Deep Learning-Based Methods for Road Extraction from High-Resolution Remote Sensing Images," Remote Sensing, vol. 16, no. 12, p. 2056, Jun. 2024, doi: 10.3390/rs16122056.
[2] J. Yang, Y. A. E. Ahmed, P. Lv, W. Zhang, and Y. An, "UBR-Net: Road Extraction from High-Resolution Remote Sensing Imagery Using Multi-Scale Attention and Cross-Residual Encoding," Canadian Journal of Remote Sensing, vol. 51, no. 1, 2025, doi: 10.1080/07038992.2025.2586320.
[3] A. Mahara, M. R. K. Khan, L. Deng, N. Rishe, W. Wang, and S. M. Sadjadi, "Automated Road Extraction from Satellite Imagery Integrating Dense Depthwise Dilated Separable Spatial Pyramid Pooling with DeepLabV3+," Applied Sciences, vol. 15, no. 12, p. 1027, Jan. 2025, doi: 10.3390/app15031027.
[4] K. Singhakhet, A. Sriaram, S. Thaiprasit, N. Tiraborisut, V. Plodprong, and T. Siriborvornratanakul, "Road Detection from Satellite Images using Semantic Segmentation," in Proc. of International Conference, Jan. 2026, pp. 37–41, doi: 10.1145/3778265.3778271.
[5] S. N. R. Karamtoth, S. Rangdal, P. Verma, K. N. Ghogale, and G. Sajeevan, "Road extraction from satellite images using Deep Learning on HPC," ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. X-5/W2-2025, pp. 301–306, 2025, doi: 10.5194/isprs-annals-X-5-W2-2025-301-2025.
[6] R. Feng, Z. Guo, X. Du, and T. Wu, "SAM2-RoadNet: Topology-Aware Multi-Scale Road Extraction from High-Resolution Remote Sensing Images," Remote Sensing, vol. 18, no. 6, p. 913, 2026, doi: 10.3390/rs18060913.
[7] J. Wang, Z. Huang, C. Ren, H. Shao, and H. Li, "Enhancing remote sensing road extraction via DS-Unet with complementary attention and surrogate gradients," Scientific Reports, vol. 16, no. 1, p. 9044, Feb. 2026, doi: 10.1038/s41598-026-39811-x.
[8] F. Sultonov, J.-H. Park, S. Yun, D.-W. Lim, and J.-M. Kang, "Mixer U-Net: An Improved Automatic Road Extraction from UAV Imagery," Appl. Sci., vol. 12, no. 4, p. 1953, Feb. 2022, doi: 10.3390/app12041953.
[9] Y.-Q. Tan, S.-H. Gao, X.-Y. Li, M.-M. Cheng, and B. Ren, "VecRoad: Point-Based Iterative Graph Exploration for Road Graphs Extraction," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Seattle, WA, USA, Jun. 2020, pp. 8910–8918, doi: 10.1109/CVPR42600.2020.00893.
[10] F. Bastani, S. He, S. Abbar, M. Alizadeh, H. Balakrishnan, S. Chawla, S. Madden, and D. DeWitt, "RoadTracer: Automatic Extraction of Road Networks from Aerial Images," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Salt Lake City, UT, USA, Jun. 2018, pp. 4720–4728, doi: 10.1109/CVPR.2018.00497.
[11] T. Sun, Z. Chen, W. Yang, and Y. Wang, "Stacked U-Nets with Multi-Output for Road Extraction," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Salt Lake City, UT, USA, Jun. 2018, pp. 202–206, doi: 10.1109/CVPRW.2018.00034.
[12] A. Abdollahi, B. Pradhan, and A. Alamri, "VNet: Road Extraction from High Resolution Aerial Imagery," arXiv preprint, May 2020, doi: 10.48550/arXiv.2005.04273.
[13] Y. Xu, Z. Xie, Y. Feng, and Z. Chen, "Road Extraction from High-Resolution Remote Sensing Imagery Using Deep Learning," Remote Sens., vol. 10, no. 9, p. 1461, Sep. 2018, doi: 10.3390/rs10091461.
[14] J. Xin, X. Zhang, Z. Zhang, and W. Fang, "Road Extraction of High-Resolution Remote Sensing Images Derived from DenseUNet," Remote Sens., vol. 11, no. 21, p. 2499, Nov. 2019, doi: 10.3390/rs11212499.
[15] S. He, F. Bastani, S. Jagwani, M. Alizadeh, H. Balakrishnan, S. Chawla, M. M. Elshrif, S. Madden, and M. A. Sadeghi, "Sat2Graph: Road Graph Extraction through Graph-Tensor Encoding," in Proc. Eur. Conf. Comput. Vis. (ECCV), Glasgow, UK, Aug. 2020, pp. 51–67, doi: 10.1007/978-3-030-58580-8_4.
[16] G. Cheng, Y. Wang, S. Xu, H. Wang, S. Xiang, and C. Pan, "Automatic Road Detection and Centerline Extraction via Cascaded End-to-End Convolutional Neural Network," IEEE Trans. Geosci. Remote Sens., vol. 55, no. 6, pp. 3322–3337, Jun. 2017, doi: 10.1109/TGRS.2017.2669341.
[17] J. Mei, R.-J. Li, W. Gao, and M.-M. Cheng, "CoANet: Connectivity Attention Network for Road Extraction from Satellite Imagery," IEEE Transactions on Image Processing, vol. 30, pp. 8540–8552, 2021, doi: 10.1109/TIP.2021.3117075.
[18] S. Xiong, C. Ma, G. Yang, Y. Song, S. Liang, and J. Feng, "Semantic segmentation of remote sensing imagery for road extraction via joint angle prediction: comparisons to deep learning," Frontiers in Earth Science, vol. 11, p. 1301281, 2023, doi: 10.3389/feart.2023.1301281.
[19] K. Zhang, A. As'arry, X. Shen, A. A. Hairuddin, M. K. Hassan, L. Zhu, and W. Qin, "DSWFNet: dual-branch fusion of spatial and wavelet features for road extraction from remote sensing images," Scientific Reports, vol. 16, no. 1, p. 3966, Dec. 2025, doi: 10.1038/s41598-025-34091-3.