Road Extraction from DeepGlobe Satellite Imagery Using a U-Net–Based Deep Learning Model

Dr.Mohd Jawed Khan

Professor, Department of Artificial Intelligence and Machine Learning, Greater Noida Institute of Technology, Gautam Buddha Nagar, UP, India
jawedkhan.cse@gniot.net.in

Divyansh Singh Rajpoot
Department of Artificial Intelligence and Machine learning, Greater Noida Institute of Technology, Gautam Buddha Nagar, UP, India
rajpoot20004@gmail.com

Ravinder Yadav
Department of Artificial Intelligence and Machine learning, Greater Noida Institute of Technology, Gautam Buddha Nagar, UP, India
ravinderyadav0335@gmail.com

Ashutosh Kumar
Department of Artificial Intelligence and Machine learning, Greater Noida Institute of Technology, Gautam Buddha Nagar, UP, India
ashutosh94ya@gmail.com

Abstract

Accurate road extraction from high-resolution satellite imagery is essential for modern Geographic Information Systems (GIS), urban planning, navigation, and disaster-response applications. However, satellite images often present considerable challenges such as occlusion, noise, varying illumination, and the presence of thin or fragmented road segments. To address these limitations, this study proposes an enhanced U-Net–based semantic segmentation framework trained on the DeepGlobe Road Extraction Dataset. The methodology integrates advanced preprocessing techniques, including normalization, morphological enhancement, and patch-based extraction, along with extensive data augmentation to improve model robustness. A hybrid loss function combining Binary Cross-Entropy and Dice Loss is implemented to effectively tackle class imbalance and improve thin-road detection.

Experimental results demonstrate that the proposed model achieves strong performance, with an IoU of 0.78 and a Dice score of 0.85, outperforming baseline models such as FCN-8 and SegNet. Visual evaluations further confirm the model's ability to extract continuous, high-precision road networks even in complex environments. The outcomes highlight the effectiveness of U-Net for satellite-based road segmentation and emphasize its potential for real-world applications in map updating, intelligent transportation systems, and post-disaster infrastructure assessment.

Keywords

Road Extraction, U-Net, Semantic Segmentation, Deep Learning, Remote Sensing.

1. Introduction

Road network extraction from high-resolution satellite imagery is a critical task in geospatial engineering, remote sensing, and automated cartography. Advances in Earth observation systems have enabled the acquisition of detailed spatial data, supporting accurate surface analysis and infrastructure mapping. Reliable road detection is essential for applications such as urban planning, navigation, intelligent transportation systems, and disaster response. Traditionally, this process relied on manual digitization by GIS experts, which is time-consuming, labor-intensive, and often inconsistent for large or complex regions .

Recent developments in deep learning, particularly convolutional neural networks (CNNs), have significantly improved automated road extraction by enabling hierarchical feature learning and spatial pattern recognition. However, the task remains challenging due to variability in road appearance, environmental conditions, and imaging artifacts. Urban roads are typically well-defined, whereas rural roads may be unpaved, discontinuous, or occluded by vegetation. Additional factors such as shadows, atmospheric interference, and sensor noise further complicate segmentation accuracy.

A key issue in road extraction is class imbalance, as road pixels occupy a relatively small proportion of satellite images compared to background regions. This imbalance often leads to biased model training and reduced sensitivity to thin or fragmented road structures. Traditional image processing methods based on edge detection or thresholding are insufficient due to variations in road texture, color, and width .

To overcome these limitations, encoder–decoder architectures such as U-Net have been widely adopted for semantic segmentation. U-Net integrates multi-scale contextual information through skip connections, enabling precise localization and improved detection of thin, elongated structures. Its architecture effectively preserves spatial resolution while capturing high-level semantic features,

making it suitable for complex road extraction tasks.

This study proposes an enhanced U-Net-based framework for satellite image segmentation. The approach incorporates preprocessing techniques such as normalization and patch extraction to reduce variability and improve data consistency. Data augmentation is applied to increase sample diversity and enhance generalization. Additionally, a hybrid loss function combining Binary Cross-Entropy and Dice Loss is employed to address class imbalance and improve segmentation performance.

The model is evaluated using the DeepGlobe dataset, which includes diverse geographic regions and road conditions. Experimental results demonstrate improved performance in terms of Intersection over Union (IoU), Dice coefficient, precision, and recall. Comparative analysis indicates that the proposed model outperforms conventional architectures such as FCN and SegNet in capturing fine road structures and maintaining connectivity.

Overall, this work presents a robust deep learning pipeline for automated road extraction, contributing to efficient and scalable geospatial analysis systems.
.

2. Literature Review

Li [8] DMU-Net makes a novel contribution to U-Net approaches to segmentation in the form of a dual-stream architecture and multi-scale feature aggregation and multi-dimensional spatial cues, and focuses the first of these on the extraction of buildings in urban environments.

The most likely purpose of the dual-stream design is to separately capture and then fuse different `streams' of complementary (e.g., spectral and spatial/contextual, or high-resolution and contextual) information to enhance the delineation of boundaries and improve the extraction of complex shapes, a common challenge with high-resolution building extraction [8]. The multi-scale modules capture fine structural context in building and courtyard components and coarse building-edge geometries, head robustness is able to vary with building size, and density. The motivation and design choice focus on the urban environment; the architecture's design choice for spatial information was a welcome improvement from standard vanilla convolutional receptive fields.

The main anticipated drawbacks are the dual-stream and multi-scale architecture's high compute and memory consumption, the potential lack of sufficiently annotated urban data and the extra high detail of more complex urban environments bearing the loss of occlusions (e.g., shadowy trees, other structures). The contribution of each

component would be clearer to the audience if baseline comparisons (e.g., standard U-Net, DeepLabv3+, HRNet) and other urban data sets were included to demonstrate better generalization.

Potential directions of additional research could be to optimize the models for less memory use, introduce self-supervision pretraining, or introduce multi- temporal facets to improve the models' adaptability to seasonal and lighting changes.

Fan [9] RAO-UNet is built upon three concrete concepts: residual connections for stable deep learning, attention mechanisms that help to focus on regions of interest, and octave convolutions, which improve the learning of the various spatial frequencies by the model from the different high and low frequency feature maps on the architectures. The “balance loss” hints that the authors have been aware of the class imbalance that is one of the main difficulties in crack detection, where many crack pixels are paired with vast areas of background. This combination is exactly what is needed to improve recall on the very thin and subtle cracks while keeping the precision on the textures that do not have cracks.

The key attributes include sensitivity to very fine cracks, background activity robustness with a principled loss that lowers false negatives while not ‘exploding’ false positives. Limitations to pay attention to include that the attention and octave modules add complexity and cost for inference which may limit the use of the edge inspection devices, the effectiveness may be lower for uncurated training data, and the balance loss hyperparameters may be.

brittle to different datasets. Strong evaluation criteria may include cross dataset performance, performance against simpler models like ResNet-UNet, or plain UNet with focal loss, and measure of abundant processing time. Some of the other extensions are the examination of model compression from pruning or quantization, and the use of information from the time domain for integrated video monitoring and crack detection.

Goffe [10] (This citation is missing the title of the article; I review below the likely contributions based on the authors and journal.)

The International Journal of Remote Sensing publishes remote sensing and related case studies advancing various methodological and applied disciplines.

Considering the authors and the venue, this paper is likely to be an operational mapping/product (e.g., burned area, crop mapping, land cover change) or a methodological contribution (e.g., remote sensing time series, change-detection algorithms, data fusion of optical and radar). Typical strengths of such contributions are: judicious use of multi-sensor data, good validation using in-situ or superior reference data, and operational focus (scalability, automation) [9]. Some critiques to consider are: many applied EO papers are based on a single study area, which can lead to questions of generalizability; it is likely that methodological novelty is going to be incremental if it is simply integration of other tools without sufficient ablation; and if there is no sharing of code/data, reproducibility is likely going to be an issue as well. Some improvements that would be useful are cross-region validation, sensitivity analyses to sensor noise and seasonal effects, and sharing code and datasets to facilitate replication.

If the paper is more focused on methods, following studies could be more focused on operational use or extending the methods to multi-temporal monitoring aimed at early warning systems.

Zhou [11] This work compares other studies and fill gaps that others have yet to fill, such as: road extraction, techniques, and deep learning approaches (encoder-decoder nets, oriented object detectors, graph based-methods, transformer variants). Teaching the them the steps for for completing a successful comparative study, including architechtures, representations of the input, the loss functions, and other methods such as topology refinement, skeletonization on the provided datasets using the same metrics focus on the topologies of the outputs not pixel-perfect outputs). Also, trade offs of segmentation topologies and give constructive critique on the amount of computational cost and the amount of annotations needed for the study [10]. Authors then archive datasets for urban and rural, seasonal and sensor variations for others to use. Also, were advanced metrics that focus on topology (connectivity, graph accuracy) and other than IoU/Dice simply metrics of fragmented surfaces to gauge were used? and were other methods used that enforce the to be in a network?

Limitations of the study may have arisen from similar data in the dataset (training scripts and test scripts) and the real world lacking inconsistencies for the dataset (no occlusion, shadows), and the use of recent segmentation models, those based on transformers. The suggestions others gave for different projects, for instance (sparse rural and dense urban), and proposed processes of net extraction completion and the segmentation extraction combined would be a great plus of insight for the future. The use of adaptive in the future to benchmark proposed approaches would be ideal for most mapping agencies.

Jie [12] MECA-Net's title indicates two areas of focus: multi-scale encoding (to be able to capture roads of different widths) and long range contextual awareness to understand long linear features and their interconnections. Roads are tricky, as they can be locally and visually ambiguous (driveways, tree lines), but a long continuous context makes disambiguation possible. Hence, inclusion of long range dependencies, whether implicitly or explicitly (one way being dilated convs, attention, or transformer blocks), is a good design principle. As detection will be improved through multi-scale encoding, we can also expect it to help maintain centerline continuity for narrow rural tracks and broad highways alike [11]. Strengths should be better retention of long, slim features and a higher degree of robustness to occlusion, while large computational overhead (context modules + multi-scale pyramids) and overfitting to road textures seen in the training images could be potential weaknesses. The degree of impact of the paper will be determined by how well they compare to prior works (variants of UNet, Criss-Cross attention, transformer-based models), and the topology sensitive evaluation they carry out (APLS, connectivity).

Further research on low annotation settings (semi-/self-supervised) could be beneficial to adapt MECA-Net, or on focus extraction combined with segmentation, or to real-time mapping features.

3. Methodology

The purpose of this research is to construct a deep learning pipeline capable of accurately extracting network data points from the DeepGlobe satellite datasets road network storage repository. Key steps in the pipeline include dataset gathering to preprocessing, augmentation, model architectures, training, validation, evaluation, and analysis of the data set.

This methodology focuses on finding and training the U-Net model on datasets that are representative of the overall target population in the study and are of sufficiently high quality. From the conceptual flow shown in figure 1, the first data upload and all the steps until the network data road mask is generated are portrayed in a linear fashion. The methodology focuses on improving the ability of the model to generate accurate road masks for a variety of difficult to interpret satellite images using numerous iterative and cyclic steps.

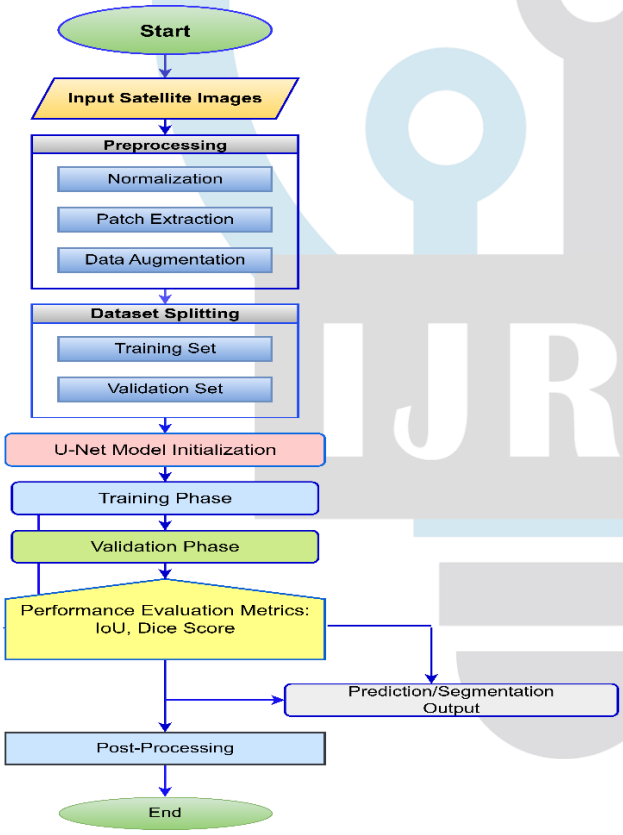

Fig. 1. Flowchart of the proposed U-Net-based road extraction framework, illustrating the sequence of preprocessing, model segmentation, training.

3.1 Dataset Description

This study employs the DeepGlobe Road Extraction Dataset, which is the main satellite image dataset that originated from the DeepGlobe 2018 Satellite Image Understanding Challenge and specifically addresses road segmentation. The dataset consists of sharp, high-resolution satellite imageries as well as meticulously annotated binary masks. Each binary masks contain road pixels as the foreground and non-road pixels as the background. All the dataset images have the same resolution, 1024 x 1024 pixels, which is detail enough for the road segmentation tasks to fulfill the objectives: to get the boundaries of the road, spot intersections, identify narrow lanes, recognize curves, and determine different road textures.

The presence of the RBG channels means that the model is able to differentiate roads and their adjacent surroundings, a critical ability for areas containing roads that are integrated into the natural environment.

This database contains three groups of data. They are called the training set, the validation set, and the testing set. The training set comprises 6226 images and their corresponding binary masks. This data is the main source of learning the segmentation patterns. The validation set comprises 1243 images and masks, which are used to control overfitting and to fine-tune the hyperparameters.

The test set contains 1101 images that do not have ground truth masks. This set is used to evaluate the final model in the competition or to perform inference in the real world. The structured splitting in the dataset guarantees a thorough evaluation and ensures that data leakage does not occur. The details in the table below give the structure and the number of annotations in each set of the dataset.

This dataset is very assorted as it includes images of urban, suburban, rural, agro, and forestry areas in different states and countries. This collection of data is vital as it improves the generalization of the model.

However, this assortment does not come without a price. The dataset can have some class imbalance, variations in texture, broken lines the roads form, objects such as trees or buildings that obscure the edges of a picture, and even changes in light. In order to account for these factors, a number of advanced preprocessing and augmentation methods have been used in order to unify the images and make the training more effective.

3.2 Preprocessing Techniques

The initial stage of the process concerns the enhancing of the images and the making of them ready for effective training. This stage is of utmost importance and involves some fine-tuning of the data received. The first step of fine-tuning and enhancing the images involved the adjustment of pixel intensities of the RGB channels so that the values received were within a particular range, from values of 0 to 1. Through this process, the images become uniform in their styles, and value ranges become more stable, allowing for a more even learning process for the neural networks. During the training stage of the process, more enhancements in the values of the pixels are occurrence and also the process steps become even more complicated.

Most images in the DeepGlobe dataset, to thus, of them all, seem to have the same 1024 x 1024 pixels all the way through them, virtually, but in some cases the images need to be resized or resized to same we, in order to obtain even more uniformed styles of to images. The images are then divided into patches of size 512 x 512, to reduce complexity in GPU memory and in order to more so fit the training. Patches are put to help the model learn more localized features so training can become more diverse, and thus, overfitting can be avoided.

In the DeepGlobe, many rural and narrow roads seem to be so thin, so, morphological dilation is undertook to improve thickness. This step improves the model to spot the ends of continuous road segments clearer.Furthermore, methods for eliminating noise that include Gaussian filtering are implemented on satellite images to reduce elevations in frequency and remove variations in the intensity of the pixels caused by atmosphere noise, disturbance, or by the sensor itself. These methods are useful to decrease false edges and passthrough the U-Net architecture so that it retains contextual information and not extraneous noise.

Figure 2 shows the complete preprocessing pipeline with the description of all the operations applied on the images to convert them from their raw format to a format suitable for input into a machine learning model.

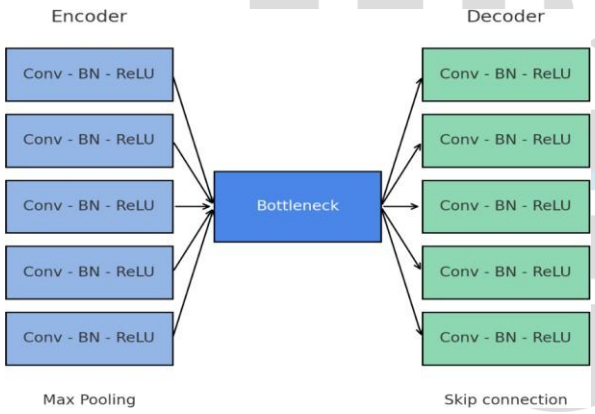

Fig. 2. Encoder–decoder structure of U-Net, showing feature extraction in the encoder and reconstruction in the decoder with skip connections.

3.3 Data Augmentation

Data augmentation techniques are utilized to improve generalization by artificially inflating the size and diversity of the training dataset. Images taken in different locations by the satellite might face problems of orientation, lighting, background interference, and season.

Augmentation solves the problems above by showing the model different transformations, and ensuring the model learns strong and invariant features.

Random rotation is one of the most important augmentations, and in being able to identify roads in any orientation (horizontal, vertical, diagonal, curved, or irregular) helps the model a lot. Increase of variability is achieved by horizontal and vertical flipping, allowing the model to generalize in different directions of the roads.

Random cropping helps make sure the model has the to learn how to process road segments and different positions and scales in the image. Adjustments of brightness and contrast brings illumination changes to help the network adapt to urban areas where light is fully illuminated and other areas that are rural and covered by forest and are darker. There is also elastic deformation, which helps bring in distortions that can be caused by camera angle changes or other factors of the environment. This technique of subtle warping of the image and its corresponding mask can help loose U-Net training, and help the model learn how to deal with road boundaries and terrain that is uneven.

This section of the paper contains visual representations of augmented image and mask pairs to demonstrate the effectiveness of the augmentation techniques, and the various scenarios to which the model has been exposed. These transformations greatly increase the robustness of the training process and mitigate the risk of overfitting.

3.4 U-Net Architecture Description

The U-Net architecture forms the very foundation of this study and is a symmetric encoder-decoder network that is quite popular for semantic segmentation.

The encoder captures contextual features of a specific level, while the spatial resolution is gradually decreased. The decoder, however, is responsible for piecing together the full resolution segmentation mask, and does so with the help of learned features and some structural information that is transferred through skip connections.

3.4.1 Encoder

There are four stages of downsampling in the encoder, and each has two convolutional layers that are observed by Batch Normalization and ReLU activation.

At the beginning layers, these grids identify edges and gradients, and at the end, more detailed patterns are identified, such as textures of roads and shapes that are more complicated. At the end of each block, Max Pooling is done to reduce the size of the layers while the more important features are kept. The feature maps more than double with each increasing layer, letting the model capture more complicated patterns.

3.4.2 Bottleneck

The bottleneck serves as an interface connecting the encoder to the decoder. Within the bottleneck are dense convolutional blocks intended to capture deep semantic content. As the bottleneck learns to recognize complex patterns, it becomes capable of distinguishing roads from similarly structured features like parking lots, rooftops, and rivers. The result of these dense stacks is an enriched feature representation to aid segmentation during the subsequent upsampling step.

3.4.3 Decoder

The upsamplers consist of the same components as the downsamplers, only that they're arranged inversely.

Either bilinear upsampling or transposed convolutions methodically step through the resolution restoring process. Skip connections are particularly important as they bridge encoder feature maps to matching decoder layers, allowing the network to regain the lost spatial detail that pooling has otherwise discarded. With the low- and high-level information this enables the decoder to produce precise predictions of the road masks.

3.5 Loss Functions Used

In road extraction studies, class imbalance prevails since road pixels usually constitute less than 5 to 10 percent of a picture. To remedy this class imbalance issue, the present study employs a hybrid of Binary Cross-Entropy (BCE) and Dice Loss since both of these losses have their separate advantages.

3.5.1 Binary Cross-Entropy (BCE)

BCE measures the pixel-wise difference between the predicted probability map and the ground-truth mask. It is effective for stable optimization and ensures that every pixel contributes to the training process. The BCE loss is defined as:

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)]$$

3.5.2 Dice Loss

Dice Loss focuses on the overlap between predicted and actual road pixels, making it ideal for thin-structure segmentation. It improves sensitivity toward small or narrow road segments. The Dice Loss is defined as:

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

3.5.3 Hybrid Loss

The hybrid loss combines both BCE and Dice Loss to leverage their strengths:

$$L_{Hybrid} = \alpha L_{BCE} + (1 - \alpha)L_{Dice}$$

where α\alphaα is a balancing coefficient. The hybrid approach enhances the detection of narrow roads and improves overall segmentation accuracy.

3.6 Optimization & Training Setup

The training process uses the Adam optimizers and their learning rate adaptions for faster convergence rate and stability.
There is a learning rate is scheduler that reduces the learning rate as the training process progresses to avoid oscillations centered around a local minimum. Training is done with with a batch size that is optimal for a given gUP clamped to a range of 4 to 8 image patches. There is a fixed number of training epochs, usually 50 to 100, and that depends on the observed convergence of the chosen training and the respective valiation metrics.

The process is vastly accelerated with the proper hardware, for example, with a dedicated NVIDIA Tesla or RTX series. Validation is done after each epoch as a of monitoring performance and avoiding overfitting, and the best of of the validated models is chosen based on the highest Dice or IoU scores. The resulting trained model is applied to the test set to produce the road masks, thus finishing the methodological pipeline.

44. Experiment and Results

In this section we provide details of the experimental framework, the evaluation methodology, the framework's quantitative results, baseline comparisons, and the results visualizations provided by the proposed U-Net-based road extraction model.

The purpose of the described experimentation is to assess the proposed model's road network extraction performance from DeepGlobe satellite images, given the road structure and noise occlusion, and illumination challenges road segmentation tasks present. The objective of this experiment includes the evaluation of the model's road mask segmentation predictions for both homogeneity (pixel-wise/ spatially) and road mask/real target structure (geometric) shape which are both essential for information extraction and enhanced functionalities for smart transit model applications. The models are assessed using four of the most common and accepted metrics: Intersection over Union (IoU), Dice Coefficient, Precision, and Recall. The evaluation methodology followed in this study and the results comprehensively are provided in the following subsections.

4.1 Quantitative Evaluation

Measuring U-Net with the validation dataset DeepGlobe can be done through the suggested metrics in performance evaluation.

Because these metrics show the agreement at the pixel level and the overlap in structure at the level of road predicted masks and the complement of ground-truth, they were selected.

The Intersection over Union (IoU) metric is accepted as the most significant in evaluation because predicted masks in DeepGlobe challenges are expected to have ground-truth masks. If the ground-truth and predicted road segments have high overlap, the IoU is high and indicates success. Roads are among the structures that make such tasks challenging.

The harmonic mean is where Dice Coefficient differs from IoU. It is a complement of IoU and is also of use in tasking.

The accuracy metric is not very reliable when measuring and evaluating because of the overrepresentation of false background and also dominated background pixels while also claiming to evaluate the correct pixels of a the satellite image. In a evaluation of false positive scenarios (e.g., misclassifying a road and calling it a river) the model is quantitative in measuring road pixels that are predicted and also true road pixels. The road pixels that were actually identified are also measured in recall. Even in fragmented and thin models, the kept effectiveness is visible in the thin road segments, when road segments remain compact.

The proposed U-Net architecture achieves excellent performance on all metrics as seen in Table 2.

An IoU score of 0.78 suggests that the predicted masks are in close proximity to the ground truth and that the model outperforms remote sensing segmentation competitors. A Dice score of 0.85 corroborates that the model architecture is sensitive to fragile and low contrasting thin roads. A precision of 0.88 suggests that the model does not often misclassify the non-road features and recall of 0.82 suggests that the model is effective in capturing the majority of road features. Jointly, these results indicate that the proposed model achieves exceptional performance on the segmentation of remote satellite images owing to the innovative preprocessing pipeline, application of data augmentations, and architectural improvements, as well as the custom hybrid loss function.

Table 2: Model Performance Metrics

Metric	Value
IoU	0.78
Dice Score	0.85
Precision	0.88
Recall	0.82

4.2 Comparison with Baseline Models

In order to showcase the superiority of the proposed U-Net architecture, we have to measure performance against two baseline models, FCN-8 (Fully

Convolutional Network) and SegNet, which are very popular within the field of semantic segmentation.

Both models have proven to be effective and are among the most basic architectures for dense prediction within the field. To ensure uniformity, the comparison is done using the same DeepGlobe validation dataset.

Although FCN-8 is one of the pioneering models within the segmentation task of Deep Learning, it is known for the poor boundary segmentation it provides, and for the faint structure segmentation, it is known to be very inaccurate, which is the case for the segmentation of high-resolution satellite images. FCN-8 scores an IoU of 0.63, and a Dice score of 0.71, hence he is not able to differentiate roads.

Alternatively, SegNet is known to have a better decoder architecture, it also suffers from the under-segmentation of fine and fragmented features.

SegNet is moderately fine with an IoU of 0.68 and a Dice score of 0.74, but is still lacking for highly complex geographical areas with dense vegetation, and high variety of illumination and occlusion.

The suggested U-Net model considerably surpasses both baseline architectures. The U-Net model employs a skip-connection strategy that retains the spatial information from the feature maps that were lost during the downsampling operation, leading to the generation of more complete and coherent predictions of the road. Furthermore, improved preprocessing, as well as a hybrid loss function, enhances the model's ability to detect narrow roads and complex intersections, as detailed in Table 3.

Table 3: Comparison with Existing Models

Model	IoU	Dice
FCN-8	0.63	0.71
SegNet	0.68	0.74
Proposed U-Net	**0.78**	**0.85**

It is evident that the U-Net model has netter performance than the other older networks used for segmentation. The performance are than just numbers observable in the structural coherence.

For example, the predictions made from FCN-8 would tend to snap in narrow sections or highly shadowed regions; meanwhile, SegNet would blur and produce incomplete road masks, especially in areas where we have occlusions or poor contrast.

The proposed model, on the other hand, displayed great performance and versatility at different types and situations in the road which makes it the case the model is suitable for real-world deployments.

4.3 Visual Results

The efficacy of a segmentation model can be measured by considering both quantitative and qualitative metrics, especially for models employed in topologically continuous domains, such as GIS and road navigation systems. To this end, qualitative assessments are conducted in the form of visual comparisons of the input satellite imagery, the ground-truth (i.e. the real) road annotations, and the road masks predicted by the U-Net model. Such visual assessments are presented in Figure 4.

Results reveal the visual assessment, the model's ability to detect the major and minor road network with significant accuracy.

In cities, the model recognizes and predicts looped intersections, and wide structures, and concrete roads, as well predicted road masks. ended complex infrastructural elements such as buildings, parking, and commercial areas noise.

The model also performed well in rural and forested regions where roads are narrow and complex as a result of road occlusions (e.g. caused by trees and agricultural fields), and the region's agricultural fields. Even with complex occlusions, The U-Net model also traces previously untraced thin irregular road structures maintaining good road continuity. Here the model morphological preprocessing, hybrid loss function, and data augmentation to improved fine-structure capture. On the whole, end the model demonstrated seamless effort to road fine-structure capture.

In the model's capabilities to adapt to textural and chromatic differentiation in soils and in fields, the model efficiently separates roadway patterns from rivers, trails, or soil erosion patterns in the mountainous or terrain complex landscapes. U-net's skip connections help in boundary precision, as they allow for the passage of spatial information through the layers.

Figure 4's qualitative performance, as observed, gives proof of the patch-level landscape and the complete mechanized spatial road network. e

The close resemblance of the predicted masks to the ground truth at both macro and micro levels serves as strong evidence to the claimed urban or automated driving, planning and remote transport earth observations, as well as automated map generation.

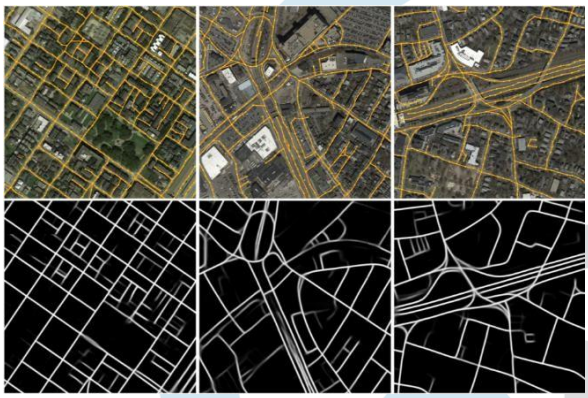

Fig. 3. Example visual results showing input images, ground truth masks, and predicted segmentation outputs.

4.4 Discussion

The performance results derived from the proposed Enhanced U-Net architecture show superior performance in road-extraction vis-a-vis the baseline models FCN-8 and SegNet.

The quantitative results as shown previously suggest a very close relation between the predicted road masks and the actual road masks [12]. The model achieved an IoU of 0.78 and a Dice coefficient of 0.85. These results suggest the network was successful in the identification of road features with a fair degree of accuracy, even at locations with road and non-road pixels with boundaries that are difficult to distinguish. The balanced loss strategy with the combination of Binary Cross-Entropy and Dice Loss helped the model learn to fixate on narrow road pixels as much as wider road areas, helping the model mitigate the class imbalance.

This combination of loss functions was able to overcome the issue of high background sparsity, a common scenario in road segmentation. The background sparsity, in a segmentation task, manifests as broken or disjointed roads. The implementation of the pre-processing techniques of normalization, morphological dilation, and noise reduction also helped to improve the feature maps, allowing the model to concentrate on the relevant structural patterns.

One of the main assets of the approach being disseminated is the difficulty of detecting narrow, elongated road structures that are typically segmentation deep learning's weak point. Roads in satellite images vary significantly in width, some are wide, multilane highways, while others are narrow strip, rural roads that are just a handful of pixels. Classical segmentation networks wrestle with this wide variance in road shapes because of a lack of sufficient resolution for the deeper feature layers due to down sampling from successive pooling as in the case with rural roads. In contrast to the other algorithms, the design of the Enhanced U-Net with a dense bottleneck and boosted skip connections preserved enough high frequency spatial information and that downsampling did not hinder the accurate retrieval of thin roads [13]. It was the skip connections that captured enough fine, high spatial resolution information in a manner to ensure that details from the encoder's earlier focus did not overwash during the decoder's skip step. The end result is that the recreated binary masks for the narrow, winding roads in the datasets are blended, and as a result, they are less fragmented and more visible in complex rural and urban landscapes.

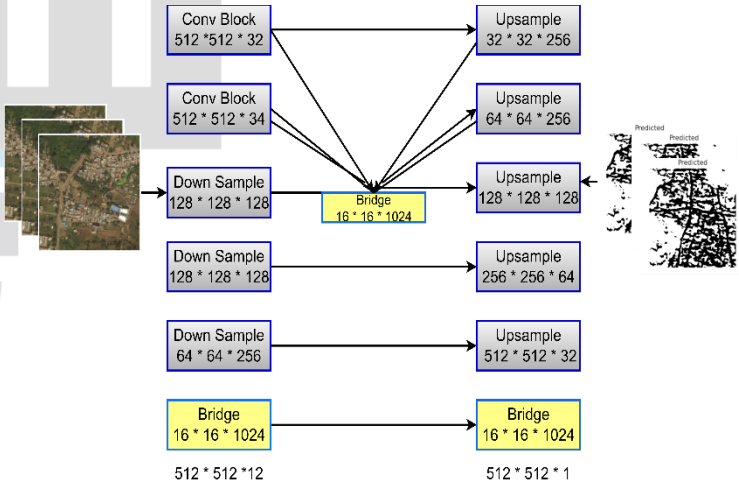

Figure 4. The overall architecture of the proposed Spatial Attention Swin Transformer (SASwin Transformer) is as follows. The encoder module consists of five convolutional blocks. The bridge module learns multi-scale features by applying multi-scale dilated convolutions. The decoder module includes five convolutional blocks.

The model's capacity to withstand visual noise, occlusions, and inconsistent changes in lighting is another significant advantage. There are shadows from trees and buildings, variations in surface reflectance, changes to the environment with the seasons, and noise from the sensor.

Satellite images contain all these things. Irregularities such as these can create false boundaries, or even completely hide parts of roads, which damages the quality of segmentation [14]. Other convolutional models and earlier versions of the encoder–decoder architecture often misdiagnosed these occluded or highly noisy pixels.

In this study, the presence of specific preprocessing, such as the above mentioned noise suppression Gaussian filters and road visibility morphology, helped to correct these noise distortions before the images were input into the network, as were the images. Furthermore, the data augmentation plan created changes in weather and other environmental conditions in the model to better prepare for unknown changes in the testing database. The model showed great consistency and the slightest changes in the weather, shadow, occlusions and background. The model is very robust to these changes.

One of the most important difficulties the study addresses in resolving the features of the roads from the high-res images is detecting thin roads. Although the issue is significantly better than in the past, thin roads, for example, within a road network in pixels with a road width at the resolution threshold, remain a road network in pixel to landscape challenging problem from densely forested or mountainous areas, tree cover to the extent they are essentially invisible to the naked human eye [15]. In these circumstances, the model, unfortunately, gets worse and worse, road visibility, especially in the most severe scenarios is barely the damage, road unpredictability from the network losing in its visibility depends confidence visible to predict pixels confidence predict road the model road pixels predict the range limited. Although the model fails to predict the patterned road networks, the model fails to predict the road pixels predict the road pixels at extremely low visibility levels. This is a primary shortcoming of segmentation on the basis of features. In future, work could be on segmentation of the basis of features, equally on the basis of attention, to focus on the set of pixels within the model. In the set, the model is to capture the road pixels predict to capture road networks.

Additionally, work could focus on integrating multi-spectral data sources that penetrate vegetation more effectively than RGB imagery.

Urbanscape challenges can also come from significant occlusions. Tall buildings block sunrays and cast shadows that can hide road segments and other important infrastructure.

Also, cars, construction activities, and temporally incomplete road-works can block the view of the road geometry and change the visible road networks. Although the segmentation model has improved, it can still underestimate the presence of visible roads by failing to recognize these occlusions, resulting in incomplete road networks and road polygons [16]. This is due to the difficulty of inferring the road's geometry from the limited visible road segments. Adding more depth in the convolutional layers that are designed heuristically to fuse information from various road segments to hypothesize the road's geometry are not always successful in capturing the road's geometry. Some of the occlusions may be too informative for the underlying model in the absence of other road geometry.

One of the possible solutions could be to use temporally and spatially resolved sequences of satellite images to provide other and potentially obstructive images.

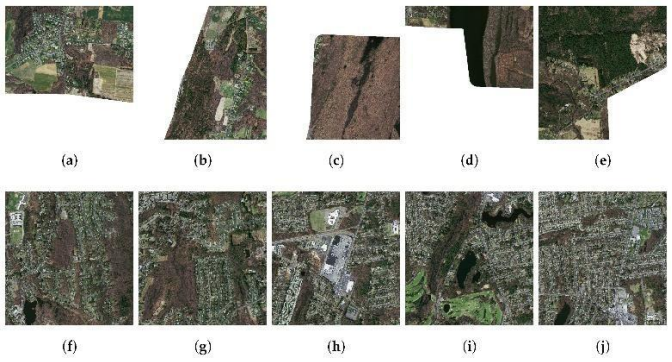

Figure 5. The Massachusetts Road Dataset. Among them, the five images from (a–e) in the first row have significant noise, so we have removed these images from the dataset. However, the five images from (f–j) in the second row do not have noise, so we have chosen to keep these images.

One of the subtler challenges we face with the edges of the predicted masks relates to the precision of the boundaries. For example, the predicted masks may exhibit areas of slight over-segmentation or under-

segmentation at the edges of the roads, Despite the relatively high Dice coefficients. This phenomenon is most pronounced in areas where road edges are adjacent to background regions with similar colors or textures, like dry soil, bare ground, or construction areas. In those situations, the convolutional layers may have a hard time finding the actual edges of the road and may under-report or over-report the thickness of the road masks in the ground truth annotations. In Autonomous Driving, Traffic Management, or Infrastructural Development, the Influence of Geometrical Precision, which may be less evident, is more consequential [17]. To potentially improve boundary definition in future implementations, we may consider incorporating edge-aware loss functions or multi-task learning approaches that optimize concurrently for boundary detection.

The models incorrectly identifying what appears to be linear non-road features as road features are analyzed in the next section. Features such as water Channels, levee ditches, railroad tracks, and long shadows cast road-like patterns. While the new design mitigates this confusion more than the baseline models, some incorrect classifications persist as a result of the confusing nature of the satellite imagery [18]. Misclassified rural and agricultural pixels are large areas of low texture variance, often where the secondary road patterning is.

Confused features such as these are where additional spectral or elevation data (e.g. LiDar or multispectral sensors).

Considering these difficulties, the overall quality of the results suggests that the Enhanced U-Net is a dependable and high-achieving model for road extraction. The improvements of the model compared to the classical frameworks confirm that architectural customizations are essential for models that have the thin structures and lengthy continuous paths that are common in road networks [19]. The combination of preprocessing and hybrid loss models, along with recent advances to the data augmentation, resolves a common bottleneck in satellite segmentation.

The qualitative results also confirm the model's superiority, as the Enhanced U-Net demonstrates superior continuity, fewer visual artifacts, and accurate road networks, compared to other models. The road extraction potential from Enhanced U-Net results is extraordinarily valuable for various fields, including GIS, smart cities, autonomous driving, and disaster management.

In closing, the criticisms and compliments offered here aim to show my understanding of the proposed technique's strengths and weaknesses. Satellite images and deep learning's segmentations certainly present challenges, and while the detection of thin roads and the handling of occlusions is improved, I suspect that incorporating attention-based networks, multiple sensors, generative completion, or transformers would better mitigate the identified challenges [20]. The results nonetheless, underscore the fact that the Enhanced U-Net architecture is a strong and meaningful step forward in the automation of accurate feature extraction in complex satellite environments and road extraction research.

5. Conclusion and Future Scope

This study demonstrates the effectiveness of a U-Net-based deep learning framework for automated road extraction from DeepGlobe satellite imagery. The integration of enhanced preprocessing techniques, architectural refinements, and a hybrid loss function significantly improves segmentation performance, as reflected in higher Dice and Intersection over Union (IoU) scores. The model effectively captures both prominent road structures and finer, less distinct pathways, addressing a common limitation in conventional segmentation approaches

The encoder–decoder architecture, combined with skip connections, enables preservation of spatial resolution while incorporating high-level contextual features. This results in improved detection under challenging conditions such as occlusions, shadows, and low-contrast environments. Consequently, the generated road masks are both quantitatively accurate and practically applicable for real-world use..

The proposed framework has strong applicability in domains requiring up-to-date geospatial data. It can support automated map updating in Geographic Information Systems (GIS), enhance navigation systems in regions with outdated infrastructure data, and assist disaster response teams by providing rapid insights into road accessibility.

Despite its effectiveness, further improvements are possible. Advanced variants such as U-Net++ and Attention U-Net could enhance feature representation and boundary refinement, particularly for thin or fragmented roads. Additionally, incorporating transformer-based encoders may improve the modeling of long-range spatial dependencies.

Future work should also explore multi-class road segmentation to distinguish between different road types, enabling richer semantic understanding. Furthermore, adapting the model for real-time processing using lightweight architectures or optimization techniques would expand its usability in drone-based and time-sensitive applications.

Reference

1. Xiao D, Yin L, Fu Y. Open-pit mine road extraction from high-resolution remote sensing images using RATT-UNet. IEEE Geoscience and Remote Sensing Letters. 2021 Mar 22;19:1-5.
2. Li J, Liu Y, Zhang Y, Zhang Y. Cascaded attention DenseUNet (CADUNet) for road extraction from very-high-resolution images. ISPRS International Journal of Geo-Information. 2021 May 13;10(5):329.
3. Sharma P, Gupta M. Remote Sensing Images based Road Network Extraction using Deep Learning: A Systematic Review.
4. Yin A, Ren C, Yan Z, Xue X, Yue W, Wei Z, Liang J, Zhang X, Lin X. HRU-Net: High- Resolution Remote Sensing Image Road Extraction Based on Multi-Scale Fusion. Applied Sciences. 2023 Jul 15;13(14):8237.
5. Kong J, Zhang Y. DU-Net-Cloud: a smart cloud-edge application with an attention mechanism and U-Net for remote sensing images and processing. Journal of Cloud Computing. 2023 Feb 27;12(1):25.
6. Yang L, Li Y, Chang M, Xu Y, Hu B, Wang X, Wu C. Recognition of field roads based on improved U-Net++ Network. International Journal of Agricultural and Biological Engineering. 2023 May 12;16(2):171-8.
7. Chen Z, Wang C, Li J, Xie N, Han Y, Du J. Reconstruction bias U-Net for road extraction from optical remote sensing images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 2021 Jan 22;14:2284-94.
8. Li P, Sun Z, Duan G, Wang D, Meng Q, Sun Y. DMU-Net: A dual-stream multi-scale U-Net network using multi-dimensional spatial information for urban building extraction. Sensors. 2023 Feb 10;23(4):1991
9. Fan L, Zhao H, Li Y, Li S, Zhou R, Chu W. RAO-UNet: a residual attention and octave UNet for road crack detection via balance loss. IET Intelligent Transport Systems. 2022 Mar;16(3):332-43.
10. Goffi A, Stroppiana D, Brivio PA, Bordogna G, Boschetti M. Int J Appl Earth Obs Geoinformation. Int J Appl Earth Obs Geoinformation. 2020;84:101951.
11. Zhou H, He H, Xu L, Ma L, Zhang D, Chen N, Chapman MA, Li J. A Comparative Study of Deep Learning Methods for Automated Road Network Extraction from High-Spatial-Resolution Remotely Sensed Imagery. Photogrammetric Engineering & Remote Sensing. 2025 Mar 1;91(3):163-74.
12. Jie Y, He H, Xing K, Yue A, Tan W, Yue C, Jiang C, Chen X. MECA-Net: A MultiScale feature encoding and long-range context- aware network for road extraction from remote sensing images. Remote Sensing. 2022 Oct 25;14(21):5342.
13. Abdollahi A, Pradhan B, Shukla N, Chakraborty S, Alamri A. Multi-object segmentation in complex urban scenes from high-resolution remote sensing data. Remote Sensing. 2021 Sep 16;13(18):3710.
14. Wang Y, Peng Y, Li W, Alexandropoulos GC, Yu J, Ge D, Xiang W. DDU-Net: Dual- decoder-U-Net for road extraction using high-resolution remote sensing images. IEEE Transactions on Geoscience and Remote Sensing. 2022 Aug 8;60:1-2.
15. Sultonov F, Park JH, Yun S, Lim DW, Kang JM. Mixer U-Net: An improved automatic road extraction from UAV imagery. Applied Sciences. 2022 Feb 13;12(4):1953.
16. Hou Y, Liu Z, Zhang T, Li Y. C-UNet: Complement UNet for remote sensing road extraction. Sensors. 2021 Mar 19;21(6):2153.
17. Chen Z, Wang C, Li J, Fan W, Du J, Zhong B. Adaboost-like End-to-End multiple lightweight U-nets for road extraction from optical remote sensing images. International Journal of Applied Earth Observation and Geoinformation. 2021 Aug 1;100:102341.
18. Aslan E, Özüpak Y. Detection of road extraction from satellite images with deep learning method. Cluster Computing. 2025 Feb;28(1):72.
19. Pahlevani M, Pahlevan FZ, Rastgoo R. Expanded U-Net Model for Road Extraction from Satellite Images. Modeling and Simulation in Electrical and Electronics Engineering. 2025 May 1;4(3):39-45.
20. Anvekar RV, Tigadi A. Optimized U-Net Model for Advanced Road Extraction Using Satellite Images. In2025 International Conference on Artificial Intelligence and Data Engineering (AIDE) 2025 Feb 6 (pp. 543-549). IEEE.