Journal of Intelligent Computing System (JICS), Vol.1, Issue. 1, pp. 53-72, (Jan 2026)
ISSN: 3139-3616 (online)

DIDRU-CAN: Dilated Inception Deep Residual U-Net with Channel Attention for Road Extraction using High-Resolution Remote Sensing Images

Palvi Sharma

Department of Computer Engineering and Applications
GLA University, Mathura, U.P, India
spalvi625@gmail.com

Abstract: Road Extraction (RE) from Remote Sensing Images (RSI) has become essential in various areas, including automatic vehicle navigation, urban planning, traffic management, and disaster management. RE from RSI is still a challenge process owing to the complex nature of the road structure and occlusions or shadows represented by nearby trees, vehicles, and buildings. A modified Dilated Inception Deep Residual U-Net Channel Attention Model (DIDRU-CAN) is embedded, combining the advantages of the Dilated Convolutional Module, the Dilated Inception Module, and the Channel Attention Mechanism (CAM). The Dilated Inception Module works effectively in capturing the multi-scale features, so that the model can extract road networks effectively. In addition, CAM provides further improves the feature selection by showing the road-related feature and suppressing irrelevant ones. CAM improves the accuracy of road extraction by utilizing skip connections that provide detailed spatial information from the decoder and high-level context information from the encoder. The Channel Attention over the skip connection provides high-level context information from the encoder and specific spatial information from the decoder to improve road extraction accuracy. In order to evaluate the performance of the proposed model, we trained the model using Massachusetts Road dataset. After the training of the model, the results shows superior performance in terms of overall accuracy, dice coefficient, and better precision and recall measures, where recall was 82.51%, precision 85.97%, overall accuracy (OA) of 97.61%, and Dice coefficient of 76.84%. Comparative studies carried out using traditional methods show that DIDRU-CAN captures complex road networks and roads that are obscured or shadowed from buildings, cars, and trees.

Keywords: *Channel Attention, Dilated Convolutional, Dilated Inception, Deep Residual U-Net, Road Extraction, Remote Sensing Images*

1. Introduction

Remote Sensing (RS) refers to the measurement and assessment of objects or regions without direct physical contact. Typically, these objects or areas can be viewed from above by airborne platforms equipped with cameras and other sensors, or from satellites. The sensors capture electromagnetic radiation that comes from different parts of the spectrum like visible, infrared, and microwave regions. Then, the radiation gathered is treated to produce mapping and imaging of the Earth's surface [1]. The transportation industry is one of the main beneficiaries of the RS technology, especially when it comes to obtaining Remote Sensing Images (RSI) traffic, related. Among other things, these pictures show the number and location of vehicles, the status of the traffic flow, and the conditions of roads such as damages, accidents, and congestion. Roads play a very important role as basic geographic information since they give necessary details about the composition and characteristics of road networks [2]. Road Extraction from RSI is key to a range of applications like emergency response planning, vehicle navigation systems, traffic management, and urban development. Besides that, the development of any country is highly dependent on the efficiency and connectivity of its road network, because well, connected roads ultimately lead to a better quality of life. It is estimated that there are around 20 million roads worldwide that have not been explored. This is why mapping uncharted and abandoned road networks is a problematic undertaking. It is difficult, costly, and time, consuming to do manual ground surveys of such road networks as they are usually geographically confined [3]. Due to obstructions by buildings, vehicles, tree shadows, and other environmental elements, road extraction from High Resolution Remote Sensing Images (HRRSI) is a very difficult task. In this regard, the automation of road extraction has been recognized as a more reasonable and convenient substitute for manual methods [4]. Road Extraction methods are categorized into two main approaches i.e., traditional method and DL based methods. Traditional techniques work on prior knowledge and superficial image characteristics

© 2026 JICS. Personal use is permitted, but republication/redistribution requires JICS permission.
see: https://synergyworldpress.com/journal/jics/about/overview

doi.org/10.67420/109319.1.1.4

such as color, texture, edges, grayscale intensity, and geometric shapes. They are examples of morphological algorithms, grayscale segmentation, and edge feature based extraction methods. They may partly succeed in extracting road information for certain images but they are limited to dynamic or complex scenarios due to environmental features like parking areas, pavement signs, building shadows, etc. [5]- [6]. Moreover, traditional techniques generally require a good knowledge of the properties, locations, and geometric features of terrestrial objects. Because of the limited features, lack of robustness, limited adaptability, and dependence on human judgment, traditional techniques are usually not suitable for real, world implementation in complicated and changing environments. In contrast DL based have received widespread recognition in scientific research as a result of the inadequacies of traditional methods. The main reasons for the use of these methods include, high productivity, adaptability, and the ability to be used in practice. Convolutional Neural Networks (CNNs) are especially powerful in semantic segmentation tasks for example road extraction from HRRSI. CNNs feature recognition capabilities in both natural and RS images have significantly contributed to the progress of automatic road extraction [7]. Despite these advances, challenges remain, particularly in precisely defining road boundaries. Feature extraction in convolutional networks can lead to the loss of important information, resulting in inaccurate segmentation [8]-[9]. Various architectures like Fully Convolutional Networks (FCN), CNN, U-Net, and Deep Residual U-Net (DRU-Net) are examples of methods that can be used for automatic road detection and delineation. Yet each method has its own drawbacks. FCN models have difficulties in detailed visualizations because of the semantic discrepancies and the limited receptive fields. CNN models are troubled in coping with the characteristics of roads and the sizes of inputs. U, Net models are the fastest but still suffer from a computational shortage when they try to capture the hierarchical features and propagate the contextual information across scales. The limitations of these methods and the applicability of extracting roads from satellite images have motivated this study to contribute to the development of more robust and accurate road extraction methods capable of addressing some of the challenging issues associated with RSI. An essential part of the work is the implementation of road segmentation techniques by embedding a Channel Attention Module (CAM) in the skip connections of Deep Residual U -Net to improve the precision and accuracy in road extraction.



Figure 1. Road Networks images and labels (a) Rural Roads (c) Small-sized towns (d) Medium-sized towns (g) Urban Roads (Ashwath, 2019)

Typically, rural roads have less crossing points and the population density is lower, as shown in Figure 1. Small to medium, sized towns are usually characterized by a smaller number of roads while urban areas have numerous junctions, intersections, parking lots, and buildings. The higher level of complexity is why the extraction of road regions from urban environments is particularly difficult. Although various methods have been utilized for road extraction from satellite images, identifying roads accurately is challenging due to the presence of buildings, trees,

and vehicles that occlude the roads. In this paper, we present the DIDRU, CAN model (Dilated Inception Deep Residual U-Net with Channel Attention) as a solution to the aforementioned issues and to the extraction of road networks that are both precise and comprehensive. The proposed model goes beyond the limitations of the conventional methods like FCN, CNN, and U-Net and at the same time makes the road extraction process more efficient. Adding dilated convolutions to the Inception modules and employing a Channel Attention Module (CAM) on the skip connections of the Deep Residual U-Net, the model achieves a number of benefits such as maintaining spatial resolution, getting the fine details, modelling the relationships between adjacent pixels, and extracting features at different levels of the hierarchy effectively. These enhancements specifically address the limitations of prior methods: CNN's restricted context capture, FCN's difficulty in preserving fine details, and U-Net's challenges in hierarchical feature extraction. Additionally, the DIDRU-CAN model improves reliability, reduces computational cost, and generalizes effectively to diverse road types and climatic conditions commonly observed in Remote Sensing Images (RSI). The work presents the following contributions: a) A DIDRU-CAN (Dilated Inception Deep Residual U-Net with Channel Attention) model is proposed to extract the roads from HRRSI using the Massachusetts Road Dataset. This model is utilized as it provides better performance and segmentation tasks in ML and computer vision sectors (image segmentation, object identification, and picture classification). b) By integrating Dilated Inception in Deep Residual U-Net, it captures the fine grained details, contextual information, and enhances the model capacity to capture hierarchical features. c) A CAM is introduced in the proposed model to highlight important information along with the channel and spatial dimension of features useful for road extraction. This mechanism highlights the essential features for the final prediction. d) A CAM is applied on "Skip Connection" to remove redundancy in the weight so that each feature map from the encoder gives more weight than creating a copy in the decoder. The objectives of the paper to carry out background research and an informative literature analysis to examine developments in road extraction techniques, with a focus on the difficulties presented by RSI. A new approach named "DIDRU-CAN" is proposed to solve the difficulties involved in extracting roads from RSI. By integrating CAM with Deep Residual Network, DIDRU-CAN improves the accuracy and resilience of road segmentation algorithms, particularly in scenarios with occlusion and shadows caused by natural and man-made structures. The novelty of DIDRU-CAN lies in its ability to prioritize relevant features while eliminating irrelevant ones, thereby improving the extraction of road characteristics in complex environments. To evaluate and verify the proposed approach using performance metrics such as dice coefficient, recall, accuracy, and precision. To compare the proposed approach with CNN, U-Net, and Deep Residual U-Net. Further, this work is categorized into different sections. Section 2 discusses the various techniques/models applied by the other researcher's in their work to extract the roads. The proposed methodology used for the model formulation and the mathematical formulation is discussed in section 3. Further, section 4 presented the experimental result formulation with its simulation. In section 5, this work is finally concluded along with its future work.

2. Background Study

In the last few decades, many of the work has been done by researchers in Road Extraction using HRRSI. For the precise and complete extraction of road networks, the researchers employed a variety of DL models. This section highlights previous studies conducted by researchers in the field of road feature extraction using HRRSI.

2.1 Road Extraction by FCN and CNN

D. Pan et al. (2021) [4] discussed the Generic Fully Convolutional Neural Network (FCNN) extract road networks from Very High Resolution Imagery (VHRI) data. The experiment was also performed training and validation on a combination of Open Street Map (OSM) geospatial data overlaid with the VHRI imagery dataset, which had spatial resolutions of both 0.3 m and 1 m. In the study, each image in the two training datasets contained a resolution of 1500 x 1500 pixels. The resulting statistics on the FCNN show that the system has a completeness rating of 94.5%, accuracy of 98.9%, and completeness rating of 93.5% high-quality road map. Q. Guo and Z. Wang (2020) [5] developed a Self-Supervised Learning Framework (SSLF) designed to extract the center of a road. In this methodology, they did not need to have a person choose the samples for training or use any optimizing procedures like eliminating non-road areas. Instead, they relied on positive sample selection techniques for finding road samples. Additionally, they introduced using Random Forests and one-class classifiers to calculate posterior probabilities for the road pixels. In summary, they based the location of the road's centerline on the Tensor Voting approach. They also performed tests with the Beijing-2 and IKONOS datasets to evaluate the scalability and robustness of the new model, demonstrating that their model achieved a completeness score of 92.05% and an accuracy score of 96.14%. X. Zhang et al. (2020) [10] proposed an FCNN (Fully Convolutional Neural Networks) for road network extraction and used to solve problems involving asymmetrical backgrounds. The FCNN included SC (Spatial Connections) by modifying the loss weights, thus providing a way to differentiate between the road and the background. When tested with the Massachusetts Road Dataset, the model achieved 72.9% accuracy. Y. Li et al. (2019) [11] utilized the framework of the VGI model to extract road regions in satellite images. The

proposed approach of extracting road region from satellite images was accomplished using satellite images of Tianjin district. Images were provided by Mapbox, Google, and Yahoo. Overall, they considered 100 images for training and testing purposes. Images have a ground resolution of around 0.48 meter per pixel and are 512 pixels by 512 pixels size. The DRLSC method is the robust segmentation of the road network that is applied to the road dataset. The curves and shapes of the roads were maintained by this method. The model obtained 80% for each precision, recall, F1 score respectively and 70 % of IOU.

J. Dai et al. (2020) [12], discussed the MSLSOH (Multi Scale Line Segment Orientation Histogram) model and sector descriptor for road tracks. The authors used three procedures to track routes in their entirety. a) The adaptive road parameter acquisition model was utilized to extract the roads' width and centerline; b) the MLSOH model was employed to forecast the road tracking direction; and c) a sector descriptor was utilized to extract the road tracking points. The suggested model provided suggestions for the road under partial shadows and retrieved the road surfaces of numerous roads obscured by the shadows of cars, buildings, and trees. The model was evaluated using the GF2 and Pleiades satellite image dataset, which had a spatial resolution of 0.8 m and 0.5 m per pixel with an image size of 2000 * 2000. The result showed that the completeness and correctness of the model were more than 98%. R. Chen et al. (2022) [13] suggested a CNN architecture to extract the road network of WUI regions using information from remote sensing images. In the WUI region, the roads are quite narrow with a large percentage of foliages in them which makes road mapping discontinuous. In that regard, the authors converted the binary classification map format into the continuous signed distance map for processing. Moreover, in the case of Massachusetts and WUI-Yajishan data sets, the recommended model achieved 64.11% and 65.92 % IOU, respectively. Y. Liu et al. (2019) [14], presented an example of ML with Multitask CNN called RoadNet that can be used for edge detection, image segmentation, and object detection from the complex urban areas using VHRI. This CNDS dataset with 224 images with a spatial resolution of 0.21 meters and having complex background scenarios of 21 cities was used for testing the proposed model. The images were divided into three main subsets: training, testing, and validation subsets, 180 images for training, 30 images for testing, and 14 images for validation purposes in order to evaluate the proposed model. The RoadNet model obtained an F1 score of 93.2 and an expected IOU score of 92.7%. Y. Wei et al. (2022) [15] developed a weakly-supervised method based on scribbles which they named the SC Road Extractor. The novel method extracts road centre-lines from satellite image tiles instead of the entire road surface. The researchers tested their system on the DeepGlobe Dataset (1665 km2 globally), the Wuhan Dataset, and the Cheng Dataset. Each of these datasets consists of 512x512 images sourced from urban, suburban and rural landscapes, with a spatial resolution of between 0.51 m and 1.23 m. To assess performance, the authors calculated the Intersection over Union (IUO) metric for the three datasets and found that their method achieved IUOs of 0.7651, 0.5158 and 0.5782 respectively. Y. Wei et al. (2020) [16], described a DL, based multi-stage Convolutional Neural Network, or CNN, framework to detect roads that were blocked by complex backgrounds like buildings and trees. In contrast to extracting the road centers and surfaces separately, the proposed model extracted the centerline and road surfaces simultaneously using the Massachusetts, Shaoxing and Cities datasets. The Massachusetts dataset had images that were 1500 x 1500 pixel images with a resolution of 1m per pixel, while the Shaoxing dataset had 532 images with 1024 x 1024 pixel images that were produced with a resolution of 0.6 m per pixel (372 images were used to train the model, and 160 images were used to test).Further, the Cities dataset used included 37 cities containing images of size 1024 x 1024 pixels at a resolution size of 60 cm by pixel. From 256 images, 192 were taken for training and another 64 images were specifically taken for testing the models. The authors reported that the multi-stage framework CNN obtained 88.2 % correctness. A. Abdollahi et al. (2020) [17], proposed a DL-based CN VNet model in order to generate a high-resolution segmentation map. In order to mitigate the issue of class imbalance in the road datasets from Massachusetts and Ottawa, the authors used the Cross-Entropy Dice Loss Function (CEDL). The authors achieved 95.89 % accuracy, which was better than the existing ones. R. Lian and L. Huang (2022) [18], discussed the Weakly Supervised road segmentation approach was using point annotations that integrated the interpretability of heuristic algorithms with the features of DCNN. Instead of labelling roads, the authors point out the existence of roads. The model was trained using the Google Earth and Massachusetts datasets, and its precision was 93.2% and 92.6%, respectively. Y. Gu et al. (2022) [19], discussed a MD-Net (Multi-Modal Network) for the extraction of roads that was appropriate for additional infrared channel RSI and increased the accuracy of road extraction by introducing infrared images. The authors achieved 62.87% accuracy using the MDNet model. S. Li (2022) [20] employed D-LinkNet50, which integrated the dilation convolution with the pre-trained LinkNet architecture. During training on the DeepGlobe dataset, the D-LinkNet50 network achieved 83.1%, 79.7%, and 81.3% accuracy, recall, and F1-score, respectively, outperforming the D-LinkNet34 network by 0.7%, 1.4%, and 1.0%. This resulted to a significant improvement in the extraction accuracy. Y. Wei, Z. Wang and M. Xu (2017) [21], proposed a RSRCNN (Road Structure Refined Convolutional Neural Network) approach to extract roads. Using the road structure's geometric data in the cross-entropy loss, a novel loss function called the "road-structure-based loss function" was introduced to train the RSRCNN. The authors used the Massachusetts Road dataset to test the proposed framework, and achieved the accuracy was 92.4% and the F1-score was 66.2%. B. Cheng et al. (2023)

[22], introduced a multi-task learning method for road edge learning, direction prediction, and road segmentation. The authors also made cascade inference using the direction estimation results and the preliminary road segmentation findings to improve the model's performance. The authors employed the DeepGlobe and SpaceNet datasets for the experimental analysis. According to the results analysis, their respective IoUs were 70.67 and 65.37%. L. Luo et al. (2022) [23], introduced BDTNet (Bi-directional Transformer) model for road extraction from the RSI. The proposed model extracted the global and local information from RSI. They used convolution operators for the extraction of local features and self-attention for global representation. Three datasets i.e. the Google Earth road dataset, the DeepGlobe dataset, and the Massachusetts road dataset were utilized to train the model. In the result analysis, they achieved 65.72%, 67.21% and 88.03% IOU respectively. C. Wang et al. (2023) [24], discussed the RSANet (Road Shape Aware Network) model for precise road extraction The authors suggested a geometric deformation estimation module for local feature extraction and an efficient strip transformer module to capture the global context. The DeepGlobe, Massachusetts, and Cheng road datasets were used for the experimental analysis. The three datasets were trained across 150 epochs with a batch size of 16. Every dataset's image size was 512 × 512. In the model evaluation, the authors obtained 78.84% precision, 86.33% recall, 82.42% F1-score, and 70.26% mIoU.

2.2 Road Extraction by U-Net and its variants

X. Lu et al. (2019) [8], presented MSMT (Multi-task Multi-Scale) DL road extraction framework for the extraction of roads as well as for centerlineThe robustness of feature extraction was enhanced by the use of U-Net as the foundational network for multi-task learning. The Massachusetts road dataset, which included 14 validation, 49 testing, and 1108 training images, was used for the experimental study. Each image size was 1500 X 1500 per pixel with a resolution of 1m per square. The proposed model achieved 52.52% accuracy. X. Li et al. (2020) [25], presented a DL study for road feature extraction. Three features of roads, i.e., pixel, edge, and region level, were extracted using the Direction Aware Attention Block (DAAB) model. DAAB enhanced the connectivity and correctness of the road where each road pixel was connected to its neighboring pixels. With a spatial resolution of 30 cm and an image size of 3000 x 3000, the authors used the dataset of four cities: Paris, Shanghai, Khartoum, and Las Vegas. M. Yuan, Z. Liu and F. Wang (2019) [26], developed U-Net based attention model to extract the road network. A partially dense-connections branch that aggregates significant feature map channels with several spatial resolutions from different encoder levels was added by the authors to the U-Net architecture. Using the Adam Optimizer, the authors applied the proposed model on the Massachusetts road dataset and obtained a learning rate of 0.0001. The authors achieved 87.56% accuracy in their experimental work. Q. Wu et al. (2021) [27], introduced the DGRN (Dense Global Residual Network), a deep learning model that reduced spatial information loss and enhanced context awareness. The proposed model attained better accuracy in GF-2 and Massachusetts Road datasets. They compared their model with the Deeplab V3, U-Net, and D-Link Net and achieved 85.00% accuracy from the proposed model. Y. Jiang et al. (2023) [28], suggested AGD-Linknet (Attention Mechanism, Gated Decoding Block, and Dilated Convolution) as a solution for problems including disjointedness, inadequate road connectivity, and incomplete highways. Three parts made up the suggested model: a) the stem block, which decreased loss and functioned as the network's first convolutional layer, b) the series-parallel combined dilated convolution and coordinate attention block, which enhanced road edge detection, and c) gated convolutional. The authors obtained an F1-score of 90.14% in an experimental analysis utilising the DeepGlobe dataset. D. Liu et al. (2022) [29], proposed a cascaded road detection network based on the Edge Sensing and Channel Attention module was proposed by D. Liu et al. (2022) [29]. Three tasks made up the suggested module: identifying the road's surfaces, identifying its borders, and determining the road's centerline. The ESM was developed to enhance the look of road regions by achieving smoother road borders. The RNDB dataset, which included 12810 training patches with 512 × 512 image sizes, was employed for the experimental study. The authors obtained 96.83% recall and 92.58% precision after the model was trained. J. Zhang et al. (2023) [30], suggested the Node Connect technique for both road topology design and node extraction from a road network. The undirected graph they used to depict the road network was written as G = (V, E), where V represented for the road's nodes and E for the node that was between the edges. Additionally, the authors employed the CNN model to forecast node connection. DeepGlobe and RoadTracer datasets were used to train the model. While training of the model the authors used a batch size of 6 and 250 epochs to train the model. Additionally, the proposed model obtained 79.95% and 83.34% precision for the DeepGlobe and RoadTracer datasets, respectively.

Y. Wang et al. (2022) [31] proposed DDU-Net, Dual Decoder U-Net for exhaustive and detailed extraction of the road network from HRRSI. The methodology DDU-Net yields accuracy and reliability in detecting small roads, significantly. Refining DDU-Net model authors integrated the channel attention block and dilated convolutional attention module in it, which have proven very effective at gathering attention sensitive features to refine the multiscale characteristics, thereby enhancing the receptive field. The proposed model was built using Massachusetts Road dataset and shows improvement over DenseUNet, DeepLabv3+ and D-LinkNet by 4%, 4.8%,

3.1% using F1 score respectively; G. Wang et al. (2023) [32] further explained that a method called Densely Connected Features (DFC-UNet) working on U-Net for Road Extraction. The authors to extract features with the help of context fusion and self-sampling learning used Dual Feature Fusion (DFF) method. The DFF method operates similar to up and down sampling techniques. Authors considered the elimination of redundancy through the substitution of DFF block for the reduction of complexity associated with high dimensional characteristics of the U-shaped structure. In addition, experimental testing on other data sets Massachusetts Road, DeepGlobe, and CHN6_CUG road dataset confirms the usability of the developed technique. Y. Zao et al. (2021) [33], introduced Richer U-Net road detecting method that combines two augmentation techniques. Initially, an EDRS (Enhanced Detail Recovery Structure) was included to identify the potential impact of convolutional processes on feature maps due to information loss. During the decoding process this structure discovers the lost features by using the convolutional layer output at the same level. To direct the network's attention to road edges the authors design an edge-focused loss function. Pixels at the edges contribute more to the loss function, increasing the amplification factor. R. Wang et al. (2023) [34], discussed a lightweight DL neural network model that integrated Multi-scale and Attention Mechanism for the extraction of road network from HRRSI. The discontinuity and edge detail loss are addressed by attention techniques. Due to the lightweight design MobileNet V2 and DeepLab V3+ was used as the backbone which reduces the training time consumption. The experiments were conducted using the road datasets from Massachusetts and Ottawa, and the accuracy rates were 98.29% and 98.92%, respectively. A. Akhtarmanesh et al. (2023) [35], suggested road extraction model using Attention –Assisted U-Net. This attention mechanism was used in the decoder section and it was trained on the patched, rotated and augmented dataset. Apart from enhancing the model's functionality, the investigation aimed to pinpoint issues that result in a decline in performance for specific image examples. Using the DeepGlobe road dataset, the experiments were carried out with 98.33% accuracy.

According to the literature review, there are three major limitations in state-of-the-art techniques for extracting roads from RSI: insufficient handling of occluded and complex backgrounds, limited scalability of the techniques across multiple diverse datasets, and a lack of a comprehensive approach. Most methods have either segmented the surface or extracted centers, meaning that the current approaches will fail to effectively combine these two aspects together. More sophisticated methods will need to be developed, such as Dilated Inception, Channel Attention Modules (CAM), and Deep Residual U-Net, to improve performance for road extraction from High Resolution Remote Sensing Imagery (HRRSI). When CAM is coupled with the skip connections of Dilated Inception Deep Residual U-Net, it enhances the model’s ability to focus on salient features for road extraction from HRRSI. The goal is to create the DIDRU-CAM model that overcomes the challenges identified in the existing literature and provides an improved method for accurately and robustly extracting roads from complicated scenes that are occluded or contain complex backgrounds. To achieve this, DIDRU-CAM continuously enhances the differentiation of road and non-road features by adding CAM to the skip connections.

3. Material and Methods

This section presents an overview of all the stages of road extraction leading up to the proposed methodology. As shown in Figure 2, a general overview of the study provides the foundational framework for this research work and consists of four primary components: collecting the data set, preparing it for analysis, training and deploying the model, and analyzing the results of the dataset.

3.1. Dataset

The dataset used for Road Extraction includes 1171 aerial images from the US state of Massachusetts, sourced from the Kaggle repository (Ashwath, 2019) [36]. Each image has a resolution of 1500 × 1500 pixels and represents an area of 2.25 square kilometres. For the purposes of training, testing, and validation, the dataset is split into 1108, 49, and 14 images, respectively. Overall, the dataset encompasses more than 2600 square kilometres and features a mix of urban, semi-rural, and rural areas. Rural areas are typically consists of few main roads and most of the area is covered by cultivated land. Semi-urban areas consists of main and link roads, less number of houses and less number of agriculture fields. Further, urban areas consists with buildings, man-made lakes, parking lots, junctions etc. The test set alone covers 110 square kilometres.

3.2. Data Pre-processing

In the context of road extraction, data pre-processing involves preparing and modifying raw data before it is used in an algorithm or machine learning model. This process typically involves various steps such as background removal, patching/segmentation and data augmentation. Data pre-processing aim is to prepare the input data for further analysis or model training by enhancing its quality and eliminating any irrelevant details. Some of the data pre –processing steps (as shown in Figure 2) are discussed below:

Background Removal: This step involves eliminating non-road elements from the image, such as vegetation, buildings, or other structures, to isolate the road network. Processing steps may be more effectively directed at the road segments by eliminating background information, which reduces computing complexity and improves accuracy.

Patching/Segmentation: To make processing easier, the image is split into smaller patches or segments after the background has been removed again. This segmentation phase improves the detection of road pixels and structures, making it possible to conduct an in-depth analysis and extraction of road information.

Data Augmentation: In the segmentation phase, data augmentation methods are used to increase the number and variability of the training dataset. Rotation, flipping, scaling, or incorporating lighting changes are some of the operations that can be utilized to improve the robustness of the model. By subjecting the model to a wider range of variability in the input data, data augmentation helps avoid overfitting and enables the model to learn more effectively from minimal training data. Data augmentation can also help address imbalances in the dataset and improve the overall efficacy of the road extraction system by creating false samples of minority classes, such as small or narrow roads.

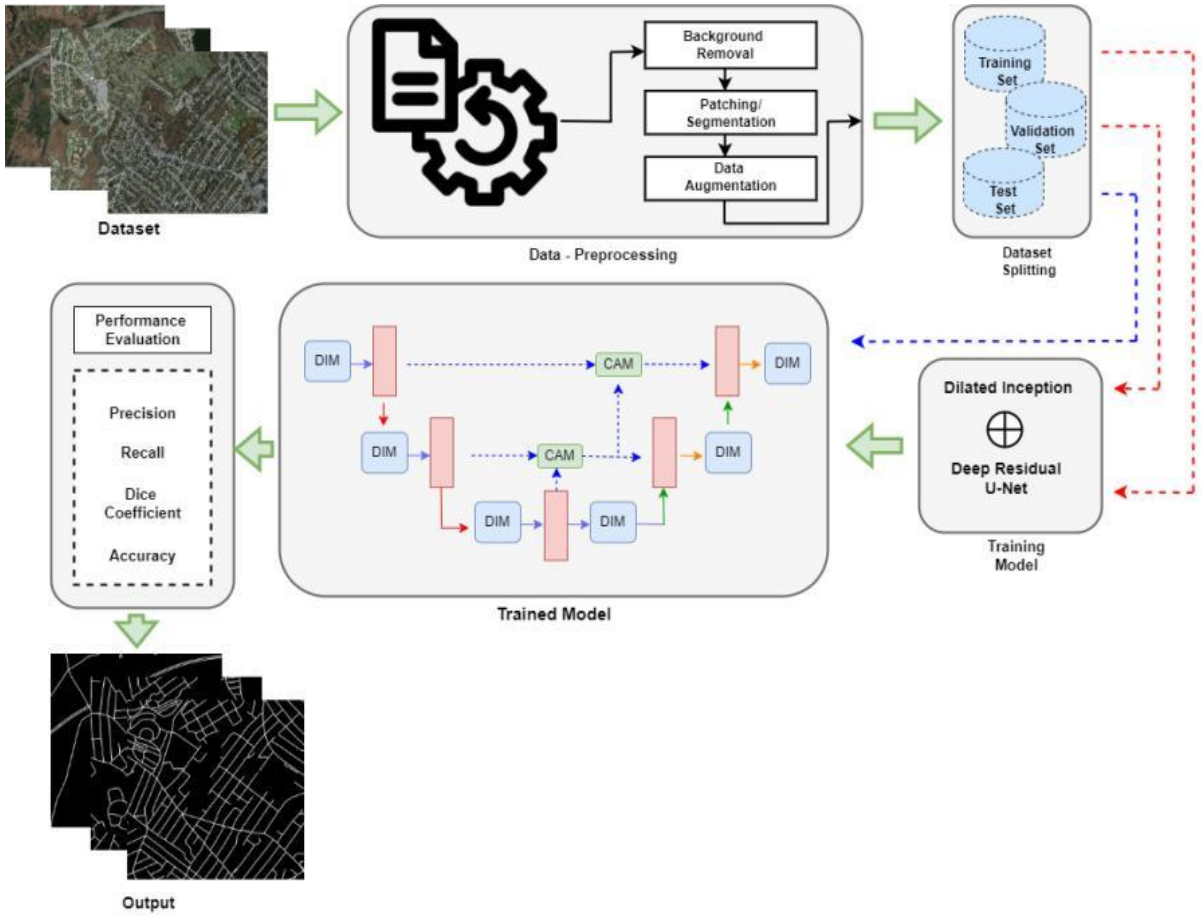


Figure 2. Steps involved in the road extraction

3.3 Proposed Methodology: DIDRU-CAN

In the road extraction from RSI, Dilated Inception Module, Deep Residual U-Net and Channel Attention are the important components that improve feature extraction and accuracy. By combining Dilated Convolution with the Inception module, the Dilated Inception Module (DIM) effectively collects contextual information at different scales in order to extract both local and global characteristics necessary for road identification. Dilated Convolutions are used to extract road details form the RSI by capturing contextual information across a large spatial area without increasing computational load or reducing resolution. The mathematical formulation of Dilated Convolution is shown below in Eq. (1):

$$Y\,[i,j] = f(\sum_{k.l} X\,[i+k, j+l] * W[k,l] + b\,) \qquad \text{Eq. (1)}$$

where, X is the input feature map, W denotes the filter, b is the bias and output is computed from Y.

$$Y\,[i,j] = f(\sum_{k.l} X\,[i + d * \ k, j + d * l] * W[k,l] + b\,) \qquad \text{Eq. (2)}$$

Eq. (2), shows the gaps between the filter elements that is identified through the dilation rate d.

Further, Dilated Inception module uses the parallel routes with different receptive field sizes to enhance the convolutional network's receptive field. The following dilated inception formulation can be used for road extraction:

$$Y = concat\ (f(X * W_1 + b_1), f\ (X * W_2\ + b_2), \ldots f(X * W_N\ +\ b_N)) \qquad \text{Eq. (3)}$$

Eq. (3), uses the symbols $*$ for the convolution operation, f for the activation function, and concat for the concatenation along the channel dimension.

With the Dilated Inception module, when it comes to extraction of road surfaces, this model is capable of recording roads at various sizes. These scales range from minute details such as road markings to more detailed contextual information, e.g. the layout of roads and surrounding areas. The inception module of the model is equipped with a dynamic convolution function, which enables it to recognize hierarchical representations of road features at different spatial dimensions. Further, the Deep Residual U-Net architecture incorporates residual connection to capture fine-grained features in road structure. This network overcomes the difficulties that is associated with deep networks and ensure more effective feature learning. Additionally, the overall performance of road extraction from RSI is improved by applying Channel Attention over the skip connections. Channel Attention selectively highlights the important features that is related to road and suppress the irrelevant ones. The integration of these elements enhances the accuracy and robustness of road extraction from complex RSI. Figure 3 depicts the DIDRU-CAN's architecture. The brief explanation of each component and the advantages of using DIDRU-CAN compared to a basic Deep Residual U-Net is discussed below:

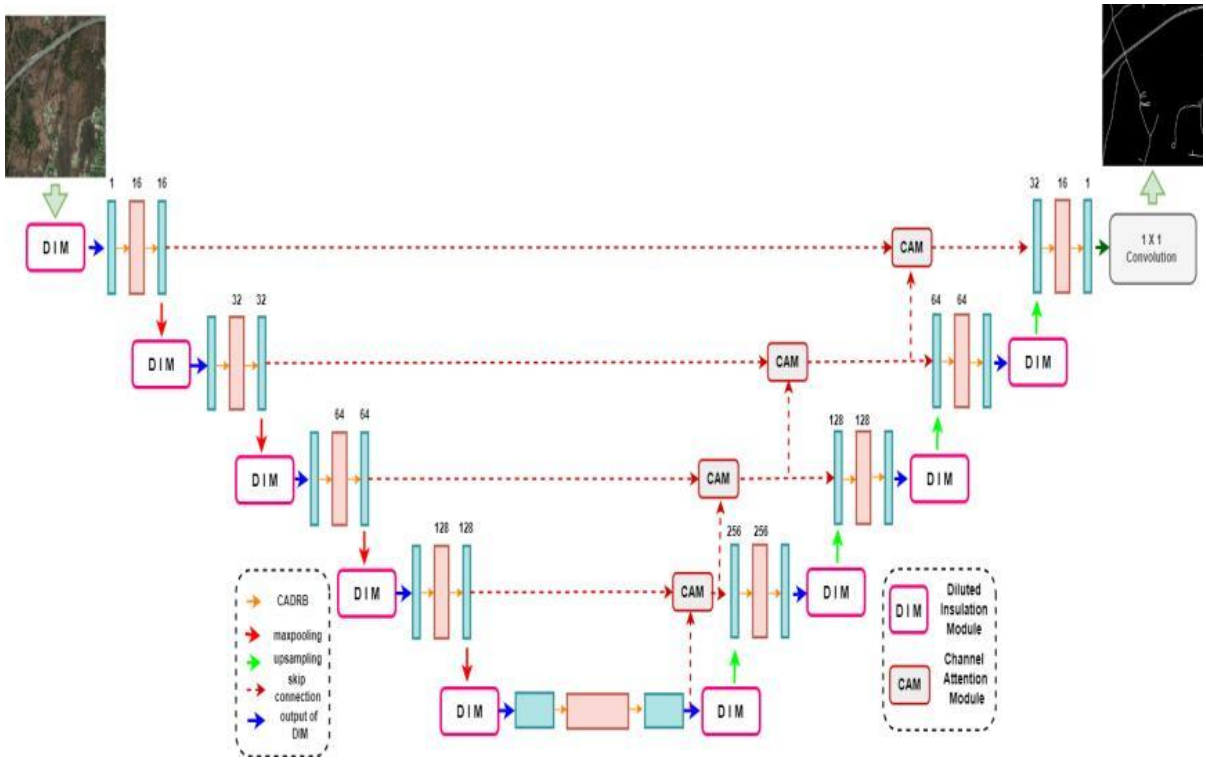


Figure 3. Proposed Architecture of the DIDRU-Net with Channel Attention Module

3.3.1 Encoder with Dilated Inception Module

Figure 3 shows the proposed Dilated Inception Deep Residual U-Net with Channel Attention (DIDRU-CAN) utilized in this study. The DIDRU-CAN consists of two main stages: the Decoder (Expansive path) and the Encoder (Contracting path), which are connected by a bridge. The encoder is positioned on the left, while the decoder is on the right, and both contain the same number of residual blocks linked by the bridge. In the Encoder, the input image, initially sized at 512 × 512, is downscaled using max-pooling layers, which reduce its dimensions by half while aligning the depth of the feature map with the number of filters in the convolutional blocks. Each

convolutional block in the encoder features two convolutional layers, with the number of filters doubling in each subsequent block. The initial block comprises of 16 filters, resulting in a map depth of 512 and dimensions of sixteen 16 x 16. In addition to improving the network's receptive field, the encoder component uses dilated convolutional blocks to preserve the spatial resolution. The network can capture more contextual information due to the gaps that dilated convolutions create between the filter elements. These dilated blocks capture local and global features such as roads. At the same time, encoder Inception modules facilitate extraction function at multiple scales. By including convolutional filters of different sizes in the same layer, encoder can capture different spatial features such as edges, textures and patterns at different scales. To identify roads in dataset, we classified pixels as either road or the background. Further, the detailed architecture of DIM and 3 x 3 dilated convolutional blocks is shown in Figure 4 and 5.

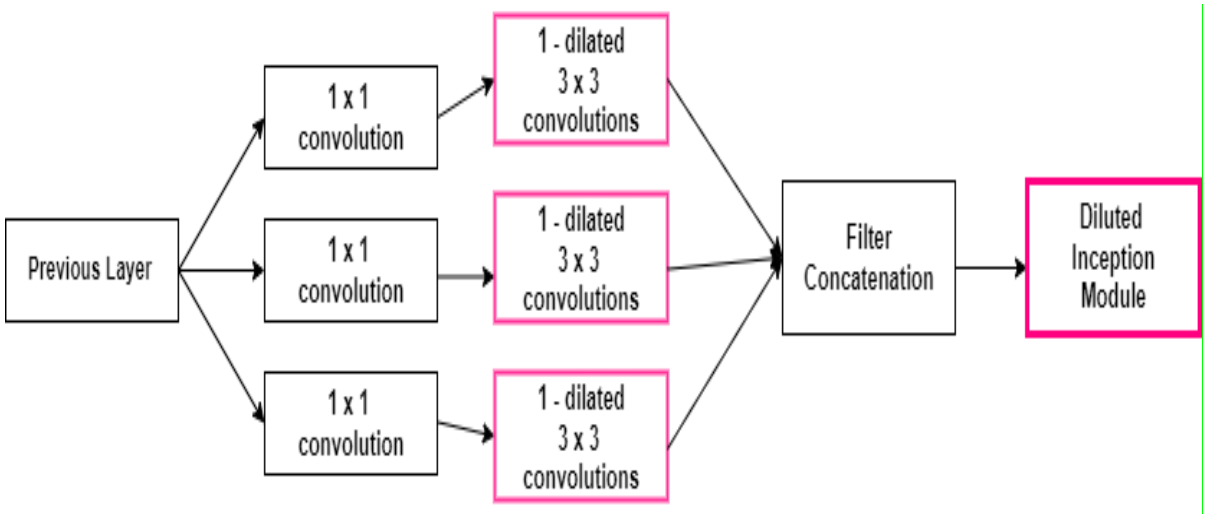


Figure 4. Dilated Inception module with 1×1 dimensional reduction convolution filters and three l-dilated convolutional filters.

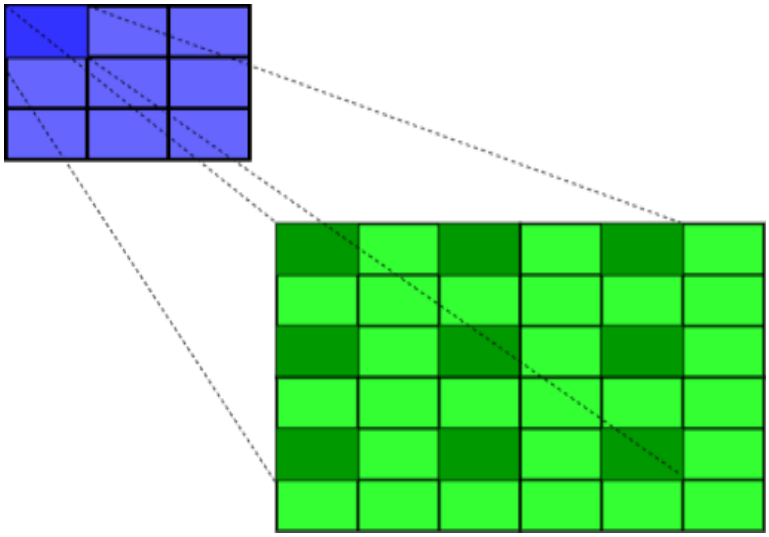


Figure 5. Dilated 3 x 3 convolution

3.3.2 Decoder with Dilated Inception Module

In the decoder part, up sampling operations are carried out in order to restore the feature maps' original spatial resolution. Additionally, skip connections are combined with the encoder's correlating feature map. This merging combines the high lever semantic information with the low-level spatial details. In addition, the decoder uses the DIM to refine feature representations and reconstruct road structures more clearly. More precise road extraction masks are generated by the Decoder by utilising multi-scale features from the Encoder. By combining DIM module in encoder and decoder section along with the CAM, the DIDRU-CAN model demonstrates improved

performance in road extraction. It effectively captures multiscale features at different scales and also improves the representation of features in RSI.

3.3.3 Channel Attention Mechanism (CAM)

This section introduces the Channel Attention Mechanism (CAM), with its complex architecture illustrated in Figure 6. The CAM is utilized over the skip connection to improve the model's focus on important features while minimizing the impact of irrelevant ones. The skip connection links the encoder and decoder parts of the Dilated Inception Deep Residual U-Net model. With the help of this skip connection encode and decoder exchange precise spatial information, also it enables the network to restore the spatial data that is lost during the down sampling. On the other hand, the CAM receives the feature maps that are obtained from the skip connections. The channel-wise information is then constructed using the input feature maps. Depending on the feature map's size, it is important to perform out global pooling in this case. Furthermore, global pooling summarises the importance of each channel's spatial extent by combining data from all the channels.

$$Y = \frac{1}{H\ X\ W} \sum_{i=1}^{H} \sum_{j=1}^{W} X\ [i, j, c] \qquad \text{Eq. (4)}$$

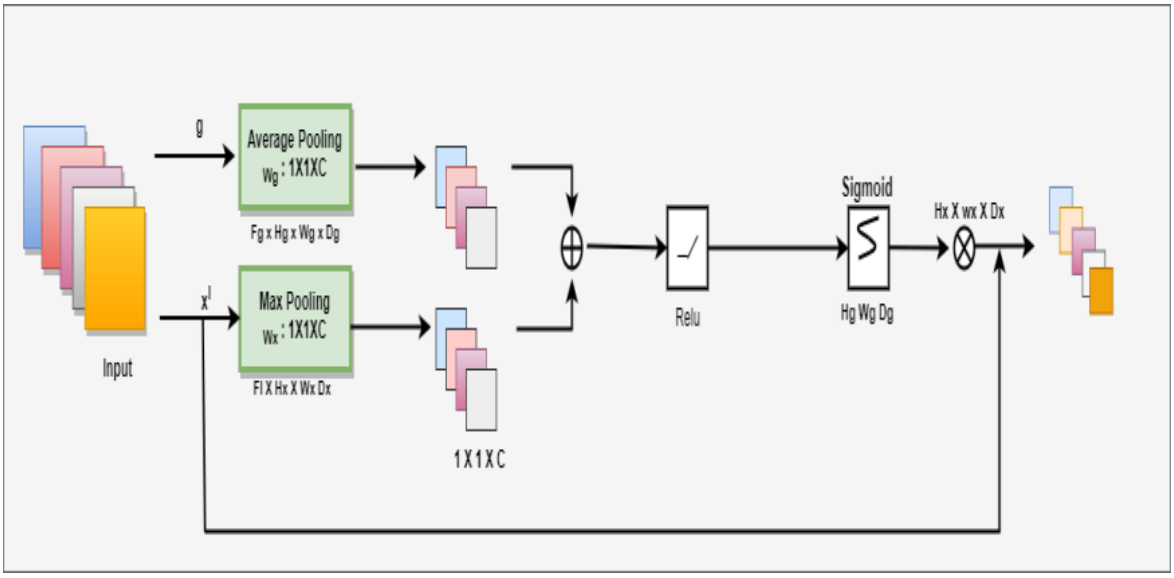
The input feature map, X, has the dimensions H X W X C in Eq. (4), where H is for height, W is for weight, and C is for the number of channels. A vector for channel-wise statistics is produced by this technique by averaging the values of each channel over all spatial areas.

$$Y = gpool\ (X) \qquad \text{Eq. (5)}$$

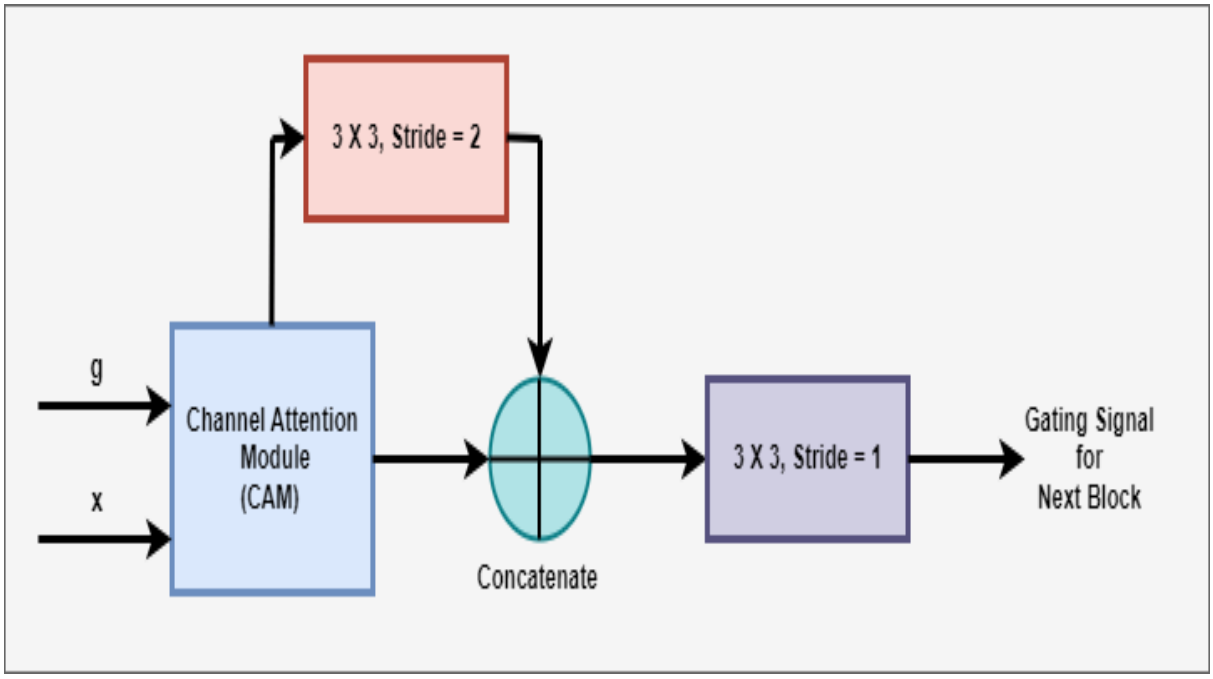
As shown in Eq. (5), Y represents the resulting vector of channel-wise statistics obtained when the global pooling procedure gpool is applied to the map X of input features. Y elements are equivalent to the statistics of a specific channel in the input feature maps.

In addition, each channel's attention weights are calculated by fully connected layers to indicate the importance of each channel for road extraction. By activating functions such as, the weights are guaranteed to be non-negative and to sum up to one. Further, by suppressing less relevant channels and strengthening information channels by means of element multiplication, estimated attention weights affect the original features maps. In the end, accurate RE from RSI is made possible by the modulation of feature maps, which are generated and subsequently processed further in the network based on trained attention rates.

By adding a CAM to Dilated Inception Deep Residual U-Net's skip connection, the model will be able to recognize road elements in RSI. For the RE, it learns to prioritize the channels by holding essential information and it will improve its segmentation accuracy and resilience. This method is based on DL architectures and attention processes, with a view to addressing the unique difficulty of extracting road data. In the end, this will enhance the ability of the model to accurately locate roadways in a complicated and overcrowded area. The detailed network structure of DIDRU-CAN is shown in Table 1.



(a)



(b)

Figure 6. (a) The CAM implementation's detailed architecture. (b) CAM position in the model

Table 1. The network structure of DIDRU-CAN

	Levels	Convolutional Layers	Filters	Stride	Output Size
Input					512×512×3
	Level 1	Conv 1 Conv 2	3×3×32 3×3×32	1 1	512×512× 32 512×512×32
Encoder	Level 2	Conv 3 Conv 4	3×3×64 3×3×64	1 1	256×256×64 256×256×64
	Level 3	Conv 5 Conv 6	3×3×128 3×3×128	1 1	128×128×128 128×128×128
	Level 4	Conv 7 Conv 8	3×3×256 3×3×256	1 1	64×64×256 64×64×256
Bridge	Level 5	Conv 9	3×3×512	1	32×32×512
	Level 6	Conv 10 Conv 11	3×3×256 3×3×256	1 1	64×64×256 64×64×256
Decoder	Level 7	Conv 12 Conv 13	3×3×128 3×3×128	1 1	128×128×128 128×128×128
	Level 8	Conv 14 Conv 15	3×3×64 3×3×64	1 1	256×256×64 256×256×64
	Level 9	Conv 16 Conv 17	3×3×32 3×3×32	1 1	512×512×32 512×512×32
Output		Conv 18	1×1×1	1	512×512×1

When comparing Deep Residual U-Net with DIDRU-CAN for the RE from RSI several differences occur. Despite its efficiency, Deep Residual U-Net may struggle to capture contextual feature and minute details that is needed for road extraction if different environmental conditions. On the other hand, DIDRU-CAN combines the dilated convolution with dilated inception and it is able to capture multi-scale contextual information as well as it handles different road widths, textures and environments efficiently. Furthermore, the proposed model captures the fine grained details in road structure and it provides better feature extraction. When these advancements combined with the Channel Attention, it produce more accurate and reliable road extraction model even in complicated environments observed in RSI. Overall DIDRU-CAN model surpasses Deep Residual U-Net in terms of accuracy and precision.

3.4 Mathematical Formulation

The mathematical formulation of the DIDRU-CAN model, which implements road extraction from HRRSI, is discussed in this section, as shown in Figure 3. In the architecture of DIDRU-CAN, Average and Max pooling functions are calculated. These pooling operations are used to pool spatial data out of a feature map, leading to two different spatial context descriptors, i.e., F^c_{avg} & F^c_{max}, where F^c is the input feature, avg is the average pooling, the maximum pooling is max, and the number of channels is c. Both average and max pooling are forwarded to the shared network to achieve the final results accurately.

$$F^c \in R^{W \times H \times C} \quad \text{Eq. (6)}$$

The Eq. (6) shows the input vector is passed to the max and average pooling on a channel-by-channel basis that generates $F^c_{max} \in R^{1 \times 1 \times C}$ and $F^c_{avg} \in R^{1 \times 1 \times C}$.

$$F^c_{max} = Max\,(F(i,j)) \quad \text{Eq. (7)}$$

$$F^c_{avg} = \frac{1}{H \times W} \sum_{u=1}^{H} \sum_{j=0}^{W} F^{c\,(i,j)} \quad \text{Eq. (8)}$$

In Eq. (7) and Eq. (8), the H, W, and C stands for height, width, and channel, respectively. F is the input feature $F^c \in R^{H \times W \times C}$. RELU activation function is used. Further, the Channel Attention uses three projected matrices $W_q \in R^{D_x \times D_q}$, $W_k \in R^{D_x \times D_k}$ and $W_v \in R^{D_x \times D_v}$ to generate the corresponding Query, Key, and Value matrix denoted by Q, K, and V. The following equations are used to calculate the attention mechanism

$$Q = XW_q \in R^{N\,x\,D_q} \quad \text{Eq. (9)}$$

$$K = XW_k \in R^{N\,x\,D_k} \quad \text{Eq. (10)}$$

$$V = XW_v \in R^{N\,x\,D_v} \quad \text{Eq. (11)}$$

In Eq. (9), Eq. (10), and Eq. (11) $X = [x_1, x_2, x_3 \ldots\ldots\ldots x_n]$ and dimensions of query, key, and vectors are computed as follows:

$$D(Q,K,V) = \rho\left(\frac{QK^T}{\sqrt{d_k}}\right) V \quad \text{Eq. (12)}$$

where, d_k is a scaling factor, and the product between Q and K^T belongs to $R^{N \times N}$. ρ a normalization function that compares the i^{th} query feature to the j^{th} query feature. The common normalization function (i.e., softmax function) is applied to Eq. (12).

$$\rho = softmax_{now}(QK^T) \quad \text{Eq. (13)}$$

In Eq. (13), each row of a matrix,$K^T(softmax_{now})$ is applied. Under the condition of the softmax normalization function, the i^{th} row of the overall result matrix obtained by the dot-product attention component as per Eq. (7) can be calculated as well.

$$D\,(Q,K,V)_i = \frac{\sum_{j=1}^{N} sim\,(q_i,k_j)v_j}{\sum_{j=1}^{N} sim\,(q_i,k_j)} \quad \text{Eq. (14)}$$

In Eq. (14), sim (q_i, k_j) calculated the similarity between q^i and k^j and sim (q_i, k_j) can be further expanded as $\emptyset\,(q_i)^T\,\emptyset\,(k_j)$. Eq. 14 is further elaborated in Eq. (15) and Eq. (16).

$$D\,(Q,K,V)_i = \frac{\sum_{j=1}^{N} \emptyset\,(q_i,k_j)v_j}{\sum_{j=1}^{N} \emptyset\,(q_i,k_j)} \quad \text{Eq. (15)}$$

$$D\,(Q,K,V)_i = \frac{\sum_{j=1}^{N} \emptyset\,(q_i)^T\,\emptyset\,(k_j)v_j}{\sum_{j=1}^{N} \emptyset\,(q_i)^T,\emptyset\,(k_j)} \quad \text{Eq. (16)}$$

Eq. (15) and Eq. (16) are further elaborated in Eq. (17)

$$D\,(Q,K,V)_i = \frac{\sum_{j=1}^{N} v_j + \left(\frac{q_i}{\|q_i\|_2}\right)^T \sum_{j=1}^{N} v_j + \left(\frac{k_j}{||k_j||_2}\right) v_j^T}{N + \left(\frac{q_i}{\|q_i\|_2}\right)^T \sum_{j=1}^{N} v_j \left(\frac{k_j}{||k_j||_2}\right)} \quad \text{Eq. (17)}$$

Eq. (17) is further vectorized, as shown in Eq. (18)

$$D\ (Q,K,V)_i\ =\ \frac{\Sigma_j V_{i,j} + \left(\frac{Q}{\|Q\|_2}\right)\left(\left(\frac{K}{\|K\|}\right)^T V\right)}{N + \left(\frac{Q}{\|Q\|_2}\right)\Sigma_j\left(\frac{K}{\|K\|_2}\right)^T_{i,j}} \quad \text{Eq. (18)}$$

3.5. Proposed Algorithm

Algorithm 1: DIDRU-CAN Model

Input: Massachusetts Road Dataset (D), strides = 1, Activation Function = RELU.

Output: Prediction of road extraction masked images.

Step 1. Importing the dataset
Step 2. Pre-process the dataset
- **a)** Original Image Size: 1500* 1500 and
- **b)** Resized to 512*512*3 as height, width, and channel due to the limitation of the GPU

Step 3. Input (training and masked images) to the DIDRU-CAN model for road feature extraction
Step 4. Encoder Block has been used to reduce the spatial dimensions in every layer using convolutional layers.
Step 5. A decoder Block has been used to increase the spatial dimensions in every layer and reduces the channels.
Step 6. CAM is applied over the skip connections to learn the weights of each channel.

$$M_c\,(F) = \sigma\big(AvgPool(F)\big) + \big(MaxPool(F)\big) = \sigma\left(W_1\big(W_0(F^c avg)\big)\right) + \quad W_{\ 1}\big(W_0(F^c\ max)\big) \quad \text{Eq. (19)}$$

W_0 and W_1 are the shared weights for both inputs, while σ stands for the sigmoid function. The Avg Pool and Max Pool are shown by Average and Max Pooling.

$$X_{CAM} = \ CAM\ (X_L, X_H) \quad \text{Eq. (20)}$$

The output of CAM is denoted by X_CAM in Eq. (20), where X_L and X_H stand for the low-level and high-level feature maps, respectively.
Step 7. Finally, Adam Optimizer is applied to compile the proposed model.

4. Experiment, Result and Analysis

This section provides a detailed discussion of the experimental setup and performance metrics. The results of the proposed model using other SOTA techniques are also shown in this section.

4.1. Experimental Setup

An NVIDIA GeForce 1080 GPU with a Tensor Flow Framework implements this model. The model's training data consists of a dataset containing 1171 1500 x 1500 pixel images that were reduced in size to 512 x 512 pixels and entered into the model. The model trains with a learning rate of 0.001 through 100 epochs using a batch size of 4. To customize the learning rate for each weight of the network respectively, Adam Optimiser was chosen as the optimization algorithm for the model. Adam Optimiser is preferred to other optimization algorithms because it can compute faster and requires fewer parameters for training. Table 2 summarizes the major parameters of the study.

Table 2: Key parameters of the model

Experimental Parameters	Parameters Value
Image Size	512 × 512
Learning Rate	0.001
Epochs	100
Batch Size	4
Optimizer	Adam Optimizer
Loss Function	Dice Loss Function

4.2 Performance Metrics

To improve the performance of a model, this algorithm uses Equations (21), (22), (23), and (24) for recall, precision, the Dice coefficient and overall accuracy (OA). In order to compute recall, precision, and the Dice coefficient, first we calculated how many true positive (TP), true negative (TN), false positive (FP) and false negative (FN) instances exist. True Positive (TP) means that the pixel in the mask has been correctly classified by the model as a road; False Negative (FN) includes pixels that are in the road mask, but were mistakenly classified as background by the model. False Positive (FP) refers to pixels in the road mask that were misclassified by the model as being road pixels instead of being classified as background; and True Negative (TN) refers to those pixels that were correctly identified by the model as background in the road mask.

Precision shows how well a classifier identifies a pixel or sample of the correct class and how many incorrectly identified samples have been predicted by the classifier as road or building predictions.

$$Precision = \frac{TP}{(TP+FP)} \quad \text{Eq. (21)}$$

Recall provides similar information as precision, with the additional benefit of providing the actual sample counts for correctly identified road or building samples as well as how many road or building samples that were incorrectly predicted as road or building predictions, respectively.

$$Recall = \frac{TP}{TP+FN} \quad \text{Eq. (22)}$$

The Dice Coefficient is often used in computer vision and deep learning to quantify the degree to which two images are similar to each other.

$$Dice\ Coefficient = 2\,X\,\frac{X \cap Y}{X \cup Y} \quad \text{Eq. (23)}$$

Overall accuracy (OA) measures the accuracy of a model by determining the total percentage of all pixels or samples in the dataset that were correctly classified or segmented by the model.

$$OA = \frac{TP+TN}{TP+FN+TN+FP} \quad \text{Eq. (24)}$$

In order to adequately evaluate road segmentation models, it is also essential to consider two more important metrics: Params, which indicates the number of parameters in a model; and FLOPs, which measures the complexity of computations being performed to create the output result of the model. In effect, these two metrics tell us how much compute was used, as well as how many parameters the model requires to produce a given output. The amount of FLOPs and Param counts for DIDR-UCAN model are 211.72 B and 11.61 M, respectively.

4.3 Results

In order to properly evaluate the proposed DIDRU-CAN model, we performed a comprehensive comparison using the Massachusetts Road dataset. We implemented several well-known RSI segmentation techniques, including CNN, U-Net, and Deep Residual U-Net, utilizing encoder-decoder architectures. Table 3 presents the evaluation metrics for different segmentation models based on this dataset. As illustrated in Figure 6, the DIDRU-CAN model is applied to the collected dataset to extract the connectivity of the road network. After processing the original images with the DIDRU-CAN model, we generated masked images and compared the results to the ground truth (masks/labels). The model was trained for 100 epochs with a learning rate set at 0.001. Based on the results achieved (shown in Figure 7), it is concluded that the model produces satisfactory results with the 85.97% precision, 82.51% recall, 76.84% dice coefficient and 97.61% OA. The DIDRU-CAN model performed well and extracted the complete and accurate roads. In Figure 7, various types of original images are used for road extraction. Figure 6 shows the original images (a), (b), (c), and (d). Figure 6 shows the binary masks of the original images in (e), (f), (g), and (h). In Figure 6, red boxes represents the missing or discontinued road extracted from the SOTA. However, while deeply analysing the particular information in the yellow boxes (as shown in Figure 6 (u), (v), (x) and (w) it has been observed that DIDRU-CAN provides a better result as attention module is induced with residual blocks. Roads in areas with similar colours as well as those obscured by tall buildings, cars, and trees can be extracted more effectively using this model. Additionally, the CNN, U-Net, and Deep Residual U-Net are compared with the proposed model. Additionally, it extracted those places that other three networks have missed as shown in Figure 6. The suggested model outperformed the SOTA in terms of overall accuracy, edge straightness, road integrity, and road network connection. Further, dice coefficient and dice loss graphs are shown in Figure 8 and 9.

Table 3: Results of Comparisons on the Massachusetts Road Dataset

Methods	Precision	Recall	Dice Coefficient	OA
CNN	74.23	60.95	40.67	77.85
U-Net	72.52	77.38	72.10	85.90
Deep Residual U-Net	74.89	80.95	73.46	90.29
DIDRU-CAN	**85.97**	**82.51**	**76.84**	**97.61**

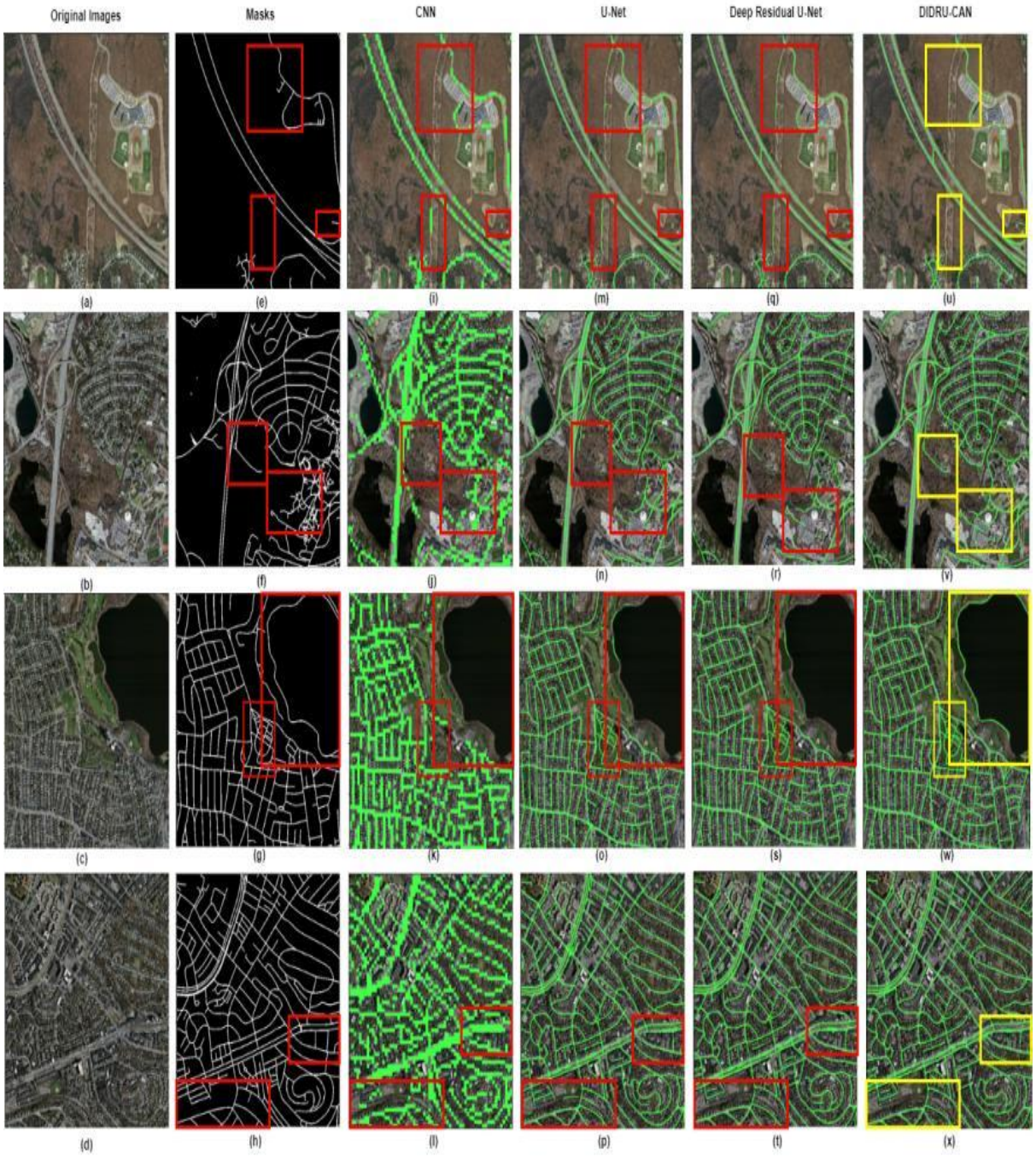


Figure 7. Comparison of road extraction result between DIDRU-CAN and traditional algorithms

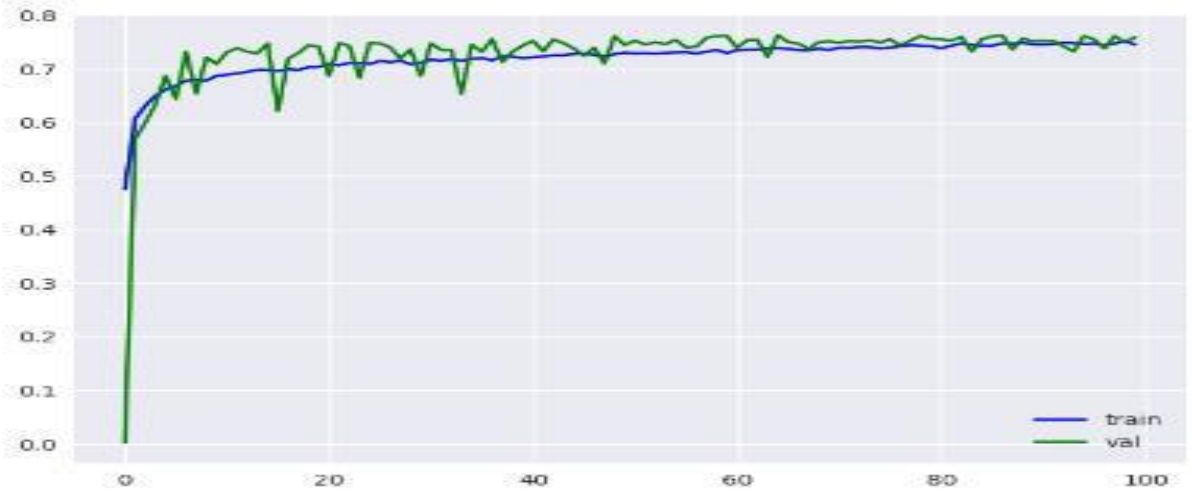


Figure 8: Dice coefficient curve during the training phase of DIDRU-CAN

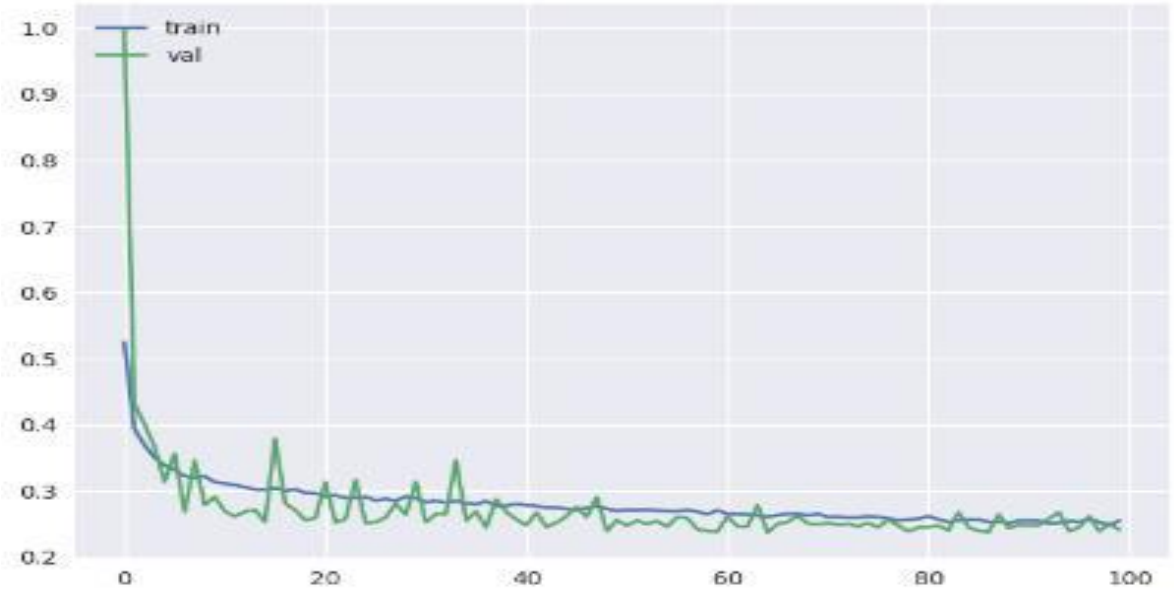


Figure 9. Dice Loss curve during the training phase of DIDRU-CAN

4.4 Comparative Analysis

The CNN, U-Net, and Deep Residual U-Net road extraction models are compared with the proposed model. Road areas are extracted using these models on the collected dataset. From the results shown in Figure 7 (i), (j), (k) and (l) it is clear that CNN model is not able to extract complete and accurate road network. The output generated from this model is not satisfactory as it provide less dice coefficient value. The dice coefficient and dice loss curve graphs are shown in Figure 10 and 11. Moreover, U-Net and Deep Residual U-Net were used to enhance the road extraction performance. The U-Net and Deep Residual U-Net findings are shown in Figure 7 (m), (n), (o), and (p) as well as Figure 7 (q), (r), (s), and (t). From the Figure 7, it is clear that both the models failed to extract complete and accurate roads where as our model provides satisfactory results. U-Net and Deep Residual UNet's dice coefficient and dice loss graphs are displayed in Figures 12, 13, 14, and 15, respectively.

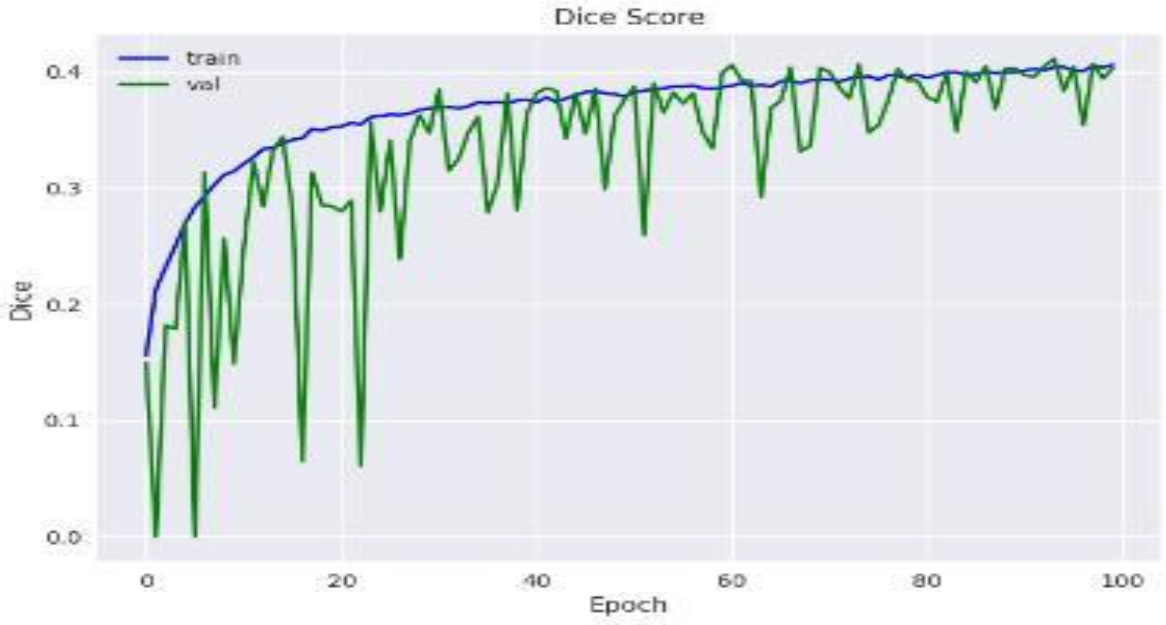


Figure 10. Dice coefficient curve during the training phase of CNN

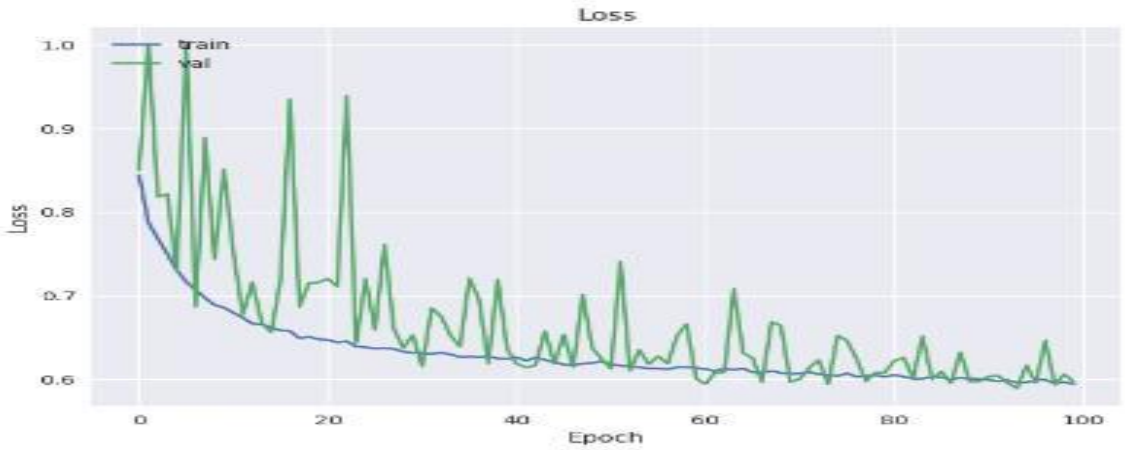


Figure 11. Dice Loss curve during the training phase of CNN

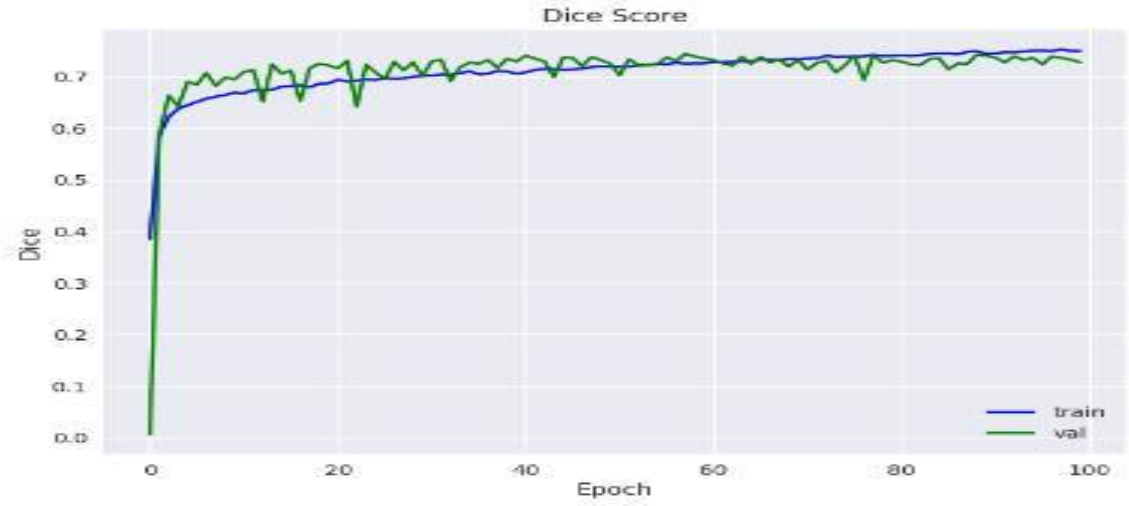


Figure 12. Dice coefficient curve during the training phase of U-Net

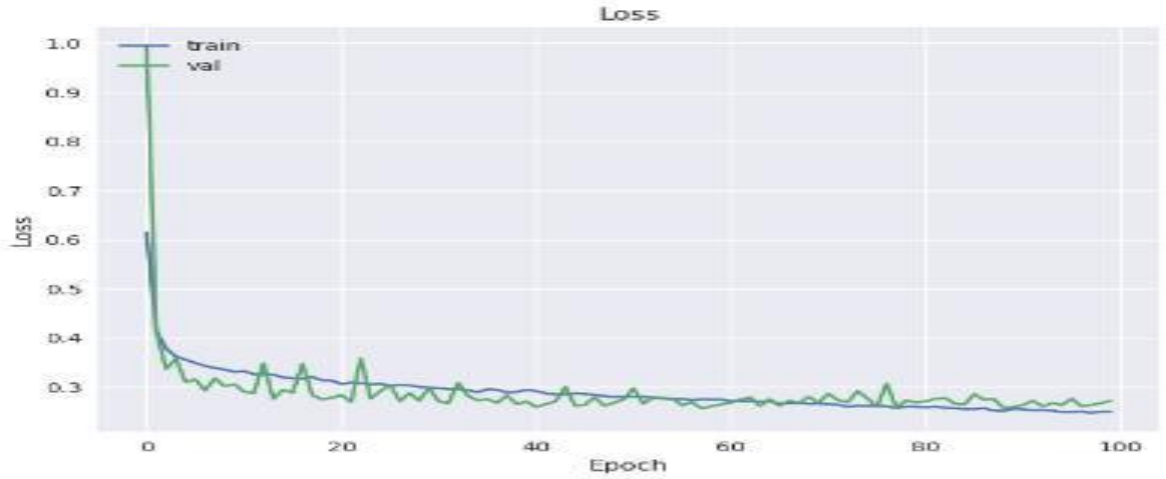


Figure 13. Dice loss curve during the training phase of U-Net

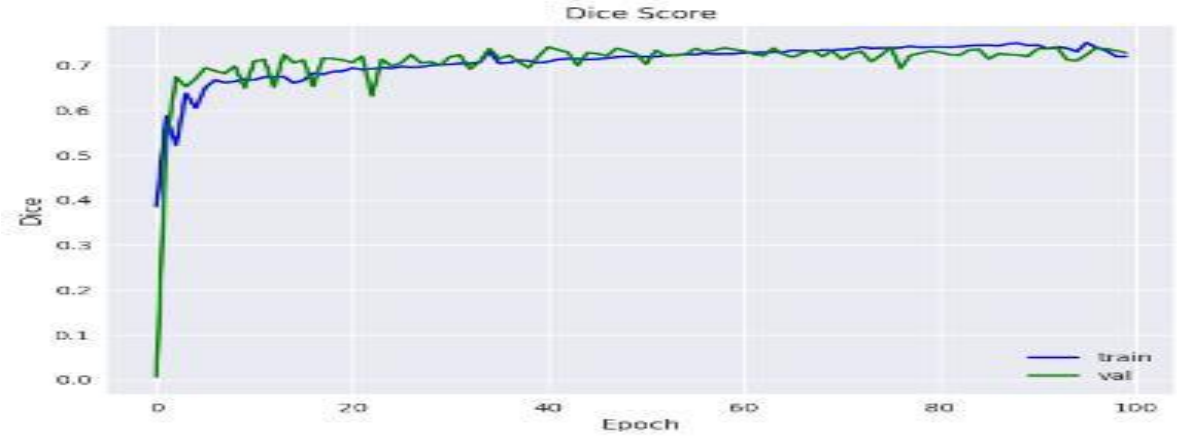


Figure 14. Dice coefficient curve during the training phase of Deep Residual U-Net

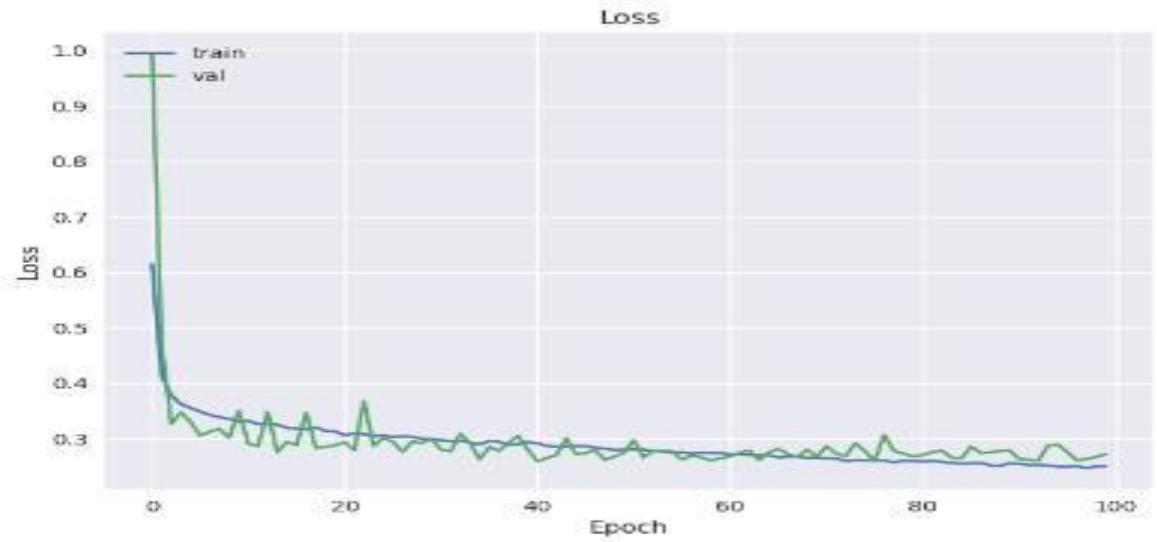


Figure 15. Dice loss curve during the training phase of Deep Residual U-Net

5. Conclusion and Future work

To extract roads from Remote Sensing Imagery (RSI), this work introduces a novel model named DIDRU-CAN, which integrates Dilated Inception, Deep Residual U-Net, and Channel Attention. This innovative model successfully identifies complex roadways in RSI with remarkable accuracy and robustness. The Deep Residual

U-Net is able to capture the contextual information at both local and global scales by using Dilated Inception. This improvement enables the network to identify complex road patterns even under difficult environmental conditions such as occlusion and shadows. Additionally, Channel Attention further refines the feature extraction by dynamically emphasizing channel-specific features, highlighting key features while suppressing unnecessary background noise. Experimental results on the Massachusetts Road dataset show that the DIDRU-CAN model outperforms the conventional models in terms of performance. Moreover, comparisons with three existing models (CNN, U-Net, and Deep Residual U-Net) on the same dataset show that the proposed model has achieved an outstanding accuracy of 97.61%, resulting in a smoother and more complete road network with less noise interference. In conclusion, this model provides a promising and effective solution to road extraction problems in RSI. However, it is worth mentioning that the DIDRU-CAN model is a heavy model that consumes a lot of resources and requires a high-performance GPU for RSI processing. In future work, model compression methods such as knowledge distillation, quantization, and pruning will be investigated to develop a lighter version of the model. This would improve the speed of processing and minimize the GPU and storage requirements without compromising accuracy. Furthermore, the use of Class Activation Mapping (CAM) on other datasets with different geographical and environmental characteristics may provide useful insights into the generalization capability of the model.

References

[1] Z. Chen *et al.*, "Road extraction in remote sensing data: A survey," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 112, no. March, 2022, doi: 10.1016/j.jag.2022.102833.

[2] A. Constantin, J. J. Ding, and Y. C. Lee, "Accurate Road Detection from Satellite Images Using Modified U-net," *2018 IEEE Asia Pacific Conf. Circuits Syst. APCCAS 2018*, pp. 423–426, 2019, doi: 10.1109/APCCAS.2018.8605652.

[3] Q. Zhu, X. Sun, Y. Zhong, and L. Zhang, "High-Resolution Remote Sensing Image Scene Understanding: A Review," *Int. Geosci. Remote Sens. Symp.*, pp. 3061–3064, 2019, doi: 10.1109/IGARSS.2019.8899293.

[4] D. Pan, M. Zhang, and B. Zhang, "A Generic FCN-Based Approach for the Road-Network Extraction from VHR Remote Sensing Images - Using OpenStreetMap as Benchmarks," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 2662–2673, 2021, doi: 10.1109/JSTARS.2021.3058347.

[5] Q. Guo and Z. Wang, "A Self-Supervised Learning Framework for Road Centerline Extraction from High-Resolution Remote Sensing Images," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 13, pp. 4451–4461, 2020, doi: 10.1109/JSTARS.2020.3014242.

[6] Y. Li, L. Xiang, C. Zhang, F. Jiao, and C. Wu, "A Guided Deep Learning Approach for Joint Road Extraction and Intersection Detection from RS Images and Taxi Trajectories," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 8008–8018, 2021, doi: 10.1109/JSTARS.2021.3102320.

[7] J. Mei, R. J. Li, W. Gao, and M. M. Cheng, "CoANet: Connectivity Attention Network for Road Extraction from Satellite Imagery," *IEEE Trans. Image Process.*, vol. 30, pp. 8540–8552, 2021, doi: 10.1109/TIP.2021.3117076.

[8] X. Lu *et al.*, "Multi-Scale and Multi-Task Deep Learning Framework for Automatic Road Extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9362–9377, 2019, doi: 10.1109/TGRS.2019.2926397.

[9] X. Qi, K. Li, P. Liu, X. Zhou, and M. Sun, "Deep Attention and Multi-Scale Networks for Accurate Remote Sensing Image Segmentation," *IEEE Access*, vol. 8, pp. 146627–146639, 2020, doi: 10.1109/ACCESS.2020.3015587.

[10] X. Zhang, W. Ma, C. Li, J. Wu, X. Tang, and L. Jiao, "Fully Convolutional Network-Based Ensemble Method for Road Extraction from Aerial Images," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 10, pp. 1777–1781, 2020, doi: 10.1109/LGRS.2019.2953523.

[11] Y. Li, L. Guo, J. Rao, L. Xu, and S. Jin, "Road segmentation based on hybrid convolutional network for high-resolution visible remote sensing image," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 4, pp. 613–617, 2019, doi: 10.1109/LGRS.2018.2878771.

[12] J. Dai, T. Zhu, Y. Wang, R. Ma, and X. Fang, "Road Extraction from High-Resolution Satellite Images Based on Multiple Descriptors," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 13, pp. 227–240, 2020, doi: 10.1109/JSTARS.2019.2955277.

[13] R. Chen, X. Li, Y. Hu, C. Wen, and L. Peng, "Road Extraction from Remote Sensing Images in Wildland-Urban Interface Areas," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022, doi: 10.1109/LGRS.2020.3028468.

[14] Y. Liu, J. Yao, X. Lu, M. Xia, X. Wang, and Y. Liu, "RoadNet: Learning to Comprehensively Analyze Road Networks in Complex Urban Scenes from High-Resolution Remotely Sensed Images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 4, pp. 2043–2056, 2019, doi: 10.1109/TGRS.2018.2870871.

[15] Y. Wei and S. Ji, "Scribble-based Weakly Supervised Deep Learning for Road Surface Extraction from Remote Sensing Images," 2020.

[16] Y. Wei, K. Zhang, and S. Ji, "Simultaneous Road Surface and Centerline Extraction from Large-Scale Remote Sensing Images Using CNN-Based Segmentation and Tracing," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8919–8931, 2020, doi: 10.1109/TGRS.2020.2991733.

[17] A. Abdollahi, B. Pradhan, and A. Alamri, "VNet: An end-to-end fully convolutional neural network for road extraction from high-resolution remote sensing data," *IEEE Access*, vol. 8, pp. 179424–179436, 2020, doi: 10.1109/ACCESS.2020.3026658.

[18] R. Lian and L. Huang, "Weakly Supervised Road Segmentation in High-Resolution Remote Sensing Images Using Point Annotations," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022, doi: 10.1109/TGRS.2021.3059088.

[19] Y. Gu, "MDNet : A Multi-modal Dual Branch Road Extraction Network Using Infrared Information," pp. 626–631, 2022.

[20] S. Li, "Multi-type road extraction and analysis of high- resolution images with D-LinkNet50," pp. 244–248, 2022.

[21] Y. Wei, Z. Wang, and M. Xu, "Road Structure Refined CNN for Road Extraction in Aerial Image," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 709–713, 2017, doi: 10.1109/LGRS.2017.2672734.

[22] B. Cheng, M. Tian, S. Jiang, W. Liu, and Y. Pang, "Multi-Task Learning and Multimodal Fusion for Road Segmentation," *IEEE Access*, vol. 11, no. January 2022, pp. 18947–18959, 2023, doi: 10.1109/ACCESS.2022.3151372.

[23] L. Luo, J. X. Wang, S. B. Chen, J. Tang, and B. Luo, "BDTNet: Road Extraction by Bi-Direction Transformer from Remote Sensing Images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022, doi: 10.1109/LGRS.2022.3183828.

[24] C. Wang *et al.*, "Towards accurate and efficient road extraction by leveraging the characteristics of road shapes," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–16, 2023, doi: 10.1109/TGRS.2023.3284478.

[25] X. Li, Y. Wang, L. Zhang, S. Liu, J. Mei, and Y. Li, "Topology-Enhanced Urban Road Extraction via a Geographic Feature-Enhanced Network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8819–8830, 2020, doi: 10.1109/TGRS.2020.2991006.

[26] M. Yuan, Z. Liu, and F. Wang, "Using the wide-range attention u-net for road segmentation," *Remote Sens. Lett.*, vol. 10, no. 5, pp. 506–515, 2019, doi: 10.1080/2150704X.2019.1574990.

[27] Q. Wu, F. Luo, P. Wu, B. Wang, H. Yang, and Y. Wu, "Automatic Road Extraction from High-Resolution Remote Sensing Images Using a Method Based on Densely Connected Spatial Feature-Enhanced Pyramid," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 3–17, 2021, doi: 10.1109/JSTARS.2020.3042816.

[28] Y. Jiang, C. Zhong, and B. Zhang, "AGD-Linknet : A Road Semantic Segmentation Model for High Resolution Remote Sensing Images Integrating Attention Mechanism , Gated Decoding Block and Dilated Convolution," *IEEE Access*, vol. 11, no. February, pp. 22585–22595, 2023, doi: 10.1109/ACCESS.2023.3253289.

[29] D. Liu, J. Zhang, K. Liu, and Y. Zhang, "Aerial Remote Sensing Image Cascaded Road Detection Network Based on Edge Sensing Module and Attention Module," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022, doi: 10.1109/LGRS.2022.3190495.

[30] J. Zhang, X. Hu, Y. Wei, and L. Zhang, "Road Topology Extraction From Satellite Imagery by Joint Learning of Nodes and Their Connectivity," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–13, 2023, doi: 10.1109/TGRS.2023.3241679.

[31] Y. Wang *et al.*, "DDU-Net: Dual-Decoder-U-Net for Road Extraction Using High-Resolution Remote Sensing Images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, doi: 10.1109/TGRS.2022.3197546.

[32] G. Wang, W. Yang, K. Ning, and J. Peng, "DFC-UNet: A U-Net Based Method For Road Extraction From Remote Sensing Images Using Densely Connected Features," *IEEE Geosci. Remote Sens. Lett.*, 2023.

[33] Y. Zao and Z. Shi, "Richer U-Net: Learning more details for road detection in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2021.

[34] R. Wang, M. Cai, and Z. Xia, "A Lightweight High-Resolution RS Image Road Extraction Method Combining Multi-scale and Attention Mechanism," *IEEE Access*, 2023.

[35] A. Akhtarmanesh, D. Abbasi-Moghadam, A. Sharifi, M. H. Yadkouri, A. Tariq, and L. Lu, "Road Extraction From Satellite Images Using Attention-Assisted UNet," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 17, pp. 1126–1136, 2024, doi: 10.1109/JSTARS.2023.3336924.

[36] Ashwath. B, "Massachusetts Roads Dataset," Aerial Images of Massachusetts to aid Machine Learning for Aerial Image Labelling, 2019 Available at: https://www.kaggle.com/datasets/balraj98/massachusetts-roads-dataset.