remote sensing

MDPI

Article

Semantic Segmentation and Edge Detection—Approach to Road Detection in Very High Resolution Satellite Images

Hamza Ghandorh [1], Wadii Boulila [2,3,*], Sharjeel Masood [4], Anis Koubaa [2], Fawad Ahmed [5] and Jawad Ahmad [6]

[1] College of Computer Science and Engineering, Taibah University, Medina 42353, Saudi Arabia; hghandorh@taibahu.edu.sa
[2] Robotics and Internet-of-Things Laboratory, Prince Sultan University, Riyadh 12435, Saudi Arabia; akoubaa@psu.edu.sa
[3] RIADI Laboratory, National School of Computer Science, University of Manouba, Manouba 2010, Tunisia
[4] HealthHub, Seol 06524, Korea; sharjeel.masood@healthhub.kr
[5] Department of Cyber Security, Pakistan Navy Engineering College, NUST, Islamabad 75350, Pakistan; fawad@pnec.nust.edu.pk
[6] School of Computing, Edinburgh Napier University, Edinburgh EH10 5DT, UK; J.Ahmad@napier.ac.uk
* Correspondence: wadii.boulila@riadi.rnu.tn

Abstract: Road detection technology plays an essential role in a variety of applications, such as urban planning, map updating, traffic monitoring and automatic vehicle navigation. Recently, there has been much development in detecting roads in high-resolution (HR) satellite images based on semantic segmentation. However, the objects being segmented in such images are of small size, and not all the information in the images is equally important when making a decision. This paper proposes a novel approach to road detection based on semantic segmentation and edge detection. Our approach aims to combine these two techniques to improve road detection, and it produces sharp-pixel segmentation maps, using the segmented masks to generate road edges. In addition, some well-known architectures, such as SegNet, used multi-scale features without refinement; thus, using attention blocks in the encoder to predict fine segmentation masks resulted in finer edges. A combination of weighted cross-entropy loss and the focal Tversky loss as the loss function is also used to deal with the highly imbalanced dataset. We conducted various experiments on two datasets describing real-world datasets covering the three largest regions in Saudi Arabia and Massachusetts. The results demonstrated that the proposed method of encoding HR feature maps effectively predicts sharp segmentation masks to facilitate accurate edge detection, even against a harsh and complicated background.

Keywords: deep learning; convolutional neural networks; 2D attention; satellite images; road segmentation; edge detection

check for updates

Citation: Ghandorh, H.; Boulila, W.; Masood, S.; Koubaa, A.; Ahmed, F.; Ahmad, J. Semantic Segmentation and Edge Detection—Approach for Road Detection in Very High Resolution Satellite Images. *Remote Sens.* **2022**, *14*, 613. https://doi.org/10.3390/rs14030613

Academic Editor: Lefei Zhang

Received: 15 November 2021
Accepted: 25 January 2022
Published: 27 January 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The rapid advancement of high-resolution (HR) satellite images has increased the need to process and extract valuable data. This process is extremely slow when performed manually; therefore, different computer vision techniques that can automatically extract helpful information from high-resolution satellite images have been developed, one of which is road segmentation. The goal of segmentation is to label each pixel in the input image [1,2]. Road segmentation is important to digital mapping automation; it has seen much attention and has developed considerably over the past few years. However, employing the road segmentation approach is challenging because of the associated background complexity, noise, shadows and occlusions. Techniques such as template matching or predefined features are high speed but offer low levels of accuracy. On the other hand, deep learning (DL) techniques that use supervised learning to train a convolutional neural network (CNN) on manually labeled data have shown promising results.

A few recent studies have led to advanced neural network architectures, while others have improved previous work remarkably. All of these studies have contributed to better segmentation performance. Deep convolutional neural networks such as UNet [3] and SegNet [4] use the famous encoder–decoder structure to segment images. The encoders in such networks encode the features by down-sampling the image gradually while increasing the receptive field in the process. The decoder then up-samples the features to recover their spatial dimensions and make the final predictions in full resolution. Long et al. [5] proposed fully convolutional networks (FCNs) for semantic segmentation. The authors used an encoder to extract features at multiple levels to make the segmentation of objects of different sizes easier. A decoder then combines the encoded features. Both UNet [3] and FCNs [5] have been the backbone of many developments.

Chen et al. [6] modified the FCN architecture, referred to as DeepLab, by using dilated-convolution layers in the last few layers to increase the receptive field while maintaining the spatial dimensions of the last few layers; maintaining spatial resolution in this way has proven to improve the segmentation of small objects. The DeepLab introduces a spatial pyramid pooling module, using dilated convolution layers with different dilation rates connected in parallel to capture context information at various ranges. The importance of this step is that it makes the segmentation of objects with a wide range of sizes more accessible.

Attention mechanisms such as the one used in [7] facilitate the segmentation process by giving higher priority to those pixels that are of more excellent value. Not all of the pixels in a feature map need equal attention; therefore, Ref. [8] used hierarchical attention maps to enable the merging of predictions from multiple scales, allowing the network to focus on the features that it thinks are more important from each scale.

Along with good network architecture, it is also essential to use a loss function that can adequately handle the imbalance of classes, especially when dealing with segmentation data in HR remote sensing images in which only a small proportion of an image is a road. The multi-scale attention mechanism in [9] generates an attention map for features at three scales.

Loss functions, such as weighted cross-entropy loss, the Tversky loss [10] and the negative log-likelihood (NLL) loss, have been used by many techniques to handle unbalanced datasets.

This paper proposes a technique that can simultaneously predict segmentation masks and the edges of objects in challenging environments, specifically roads in satellite images. It first generates road segmentation masks and then uses these segmented masks to create the road's edges. The main contributions in this research are as follows:

1. This study designs an encoder with a large receptive field, meaning that it can adequately segment large objects and encode features in full resolution. This step is essential, as a high resolution is needed to produce fine segmentation masks.
2. The study then uses these features to generate fine segmentation masks, which are then used to create road edges.
3. The study also implements and tests the combination of weighted cross-entropy and the Tversky loss functions, training the network to handle highly imbalanced data.

The rest of the paper is organized as follows. Section 2 provides a brief discussion on relevant related works. Section 3 summarizes essential concepts that will be used in this paper. The proposed research methodology is detailed in Sections 4 and 5. Experimental results using two different datasets are presented in Section 6. Discussion is provided in Section 7. Finally, Section 8 concludes the paper.

2. Literature Review

Several research studies have been conducted to improve the literature's road detection/extraction efficiency and performance. This section introduces recent state-of-the-art approaches for road detection and extraction.

Road detection is critical in many applications, including infrastructure planning and traffic routing software, emergency management and urban planning. Automatic road network extraction techniques have significantly increased the extraction rate of

road networks. Over the last few decades, researchers have proposed numerous new or improved image segmentation methods for road extraction.

Traditional and deep learning-based methods are the two broad categories of these methods. Traditional road extraction methods are primarily based on the assumption that the grayscale value inside the road is relatively consistent and contrasts with the surrounding objects, such as trees and buildings, ensuring road area distinguishability and severability. In contrast to traditional methods, deep learning-based methods rely on advances in feature learning and parameter sharing, which can be used to achieve automatic and efficient road extraction [11,12]. Deep learning techniques have made significant advancements in image object segmentation. The efficiency of road extraction can be scaled according to processing power and the size of the training data set [13].

To gather contextual information from different resolutions, the Pyramid Scene Parsing (PSP) Network [14] introduces a multi-path feature extraction module. This method of gathering contextual information aids in the segmentation of objects of various sizes. Henry et al. [15] presented an evaluation of fully convolutional neural networks (FCNNs) for road segmentation in high-resolution synthetic aperture radar images. To extract roads, the authors modified FCNNs by including a tolerance rule for spatially small errors. Deeplabv3+ was modified with a class-weighted mean-squared error loss to achieve a 44% IoU across the test data.

Xin et al. [16] developed a CNN-based approach, DenseUNet, to extract the road network from RS images with few parameters and robust characteristics. The model combines dense connection and U-Net to solve the tree and shadow occlusion problems and emphasize foreground pixels. Experiments were conducted on two datasets of high-resolution images and compared with three classical semantic segmentation methods. Chen et al. [17] suggested a DL-based model using a dense feature pyramid network to consider the particularity and complexity of instance segmentation of roads. In this work, shallow feature channels are concatenated with deep feature channels to use in-depth features for the shallow feature maps with high-resolution images. The authors suggested using the focal loss function to compute mask loss in the DL model better to consider the hard-classified samples with the less-pixel foreground. The experiments depicted good performances of the proposed method in improving the instance segmentation of road marking compared to the state-of-the-art methods. Instead of using pooling layers, which increase the receptive fields at the cost of reducing the spatial dimensions (see, for example, [18–22]), authors have maintained the feature maps in full resolution throughout the network to segment small objects and properly segment fine edges. Boulila et al. [23] used long short-term memory units to predict urban expansions. The images are first segmented using unsupervised learning to segment different land areas before segmenting them further using a convolutional LSTM (ConvLSTM) approach. The technique in [24] uses a CNN with a fully connected layer to classify satellite images into the water, soil, road, vegetation and urban classes. Alkhelaiwi et al. [25] used a CNN to encode satellite images, thereby reducing the overall computational costs. To segment the road surface from the remote sensing images, MRENet [26] used an encoder–decoder architecture similar to UNet [3] and a PSP module with four subregions between the encoder and the decoder modules. Brewer et al. [27] adopted a transfer learning approach based on CNN architecture trained on data collected in the United States and then fine-tuned data collected from Nigeria. Brewer et al. [27] developed an Android application to collect road data and used it as input to test several CNN architectures. The proposed approach achieved an accuracy of 80%, with 99.4% of predictions falling within the actual or adjacent class. The authors demonstrated that by tailoring the United States model based on the Nigerian data, they could reach an accuracy of 94% in predicting the quality of Nigerian roads. Li et al. [7] generated attention maps in each resolution before passing them on to the decoder. They introduce two attention mechanisms: the kernel attention mechanism (KAM) and the channel attention mechanism (CAM). The purpose of these two mechanisms is to reduce the overall computational

complexity. Instead of generating a single 3D attention map, Ref. [28] used a channel attention module and a spatial attention module separately.

The channel attention module outputs a tensor of shape $1 \times 1 \times C$, where C represents the number of channels, while the spatial attention module outputs a 2D attention map of the shape of the input. HED-UNet [8] uses an encoder–decoder network similar to that employed by UNet. Their main contribution is combining the coastline's segmentation and edge detection. They have used hierarchical attention maps, allowing the network to give adequate attention to the coast. Edge-FCN [29] combines an FCN with a Holistically Nested Edge detection method [30] to predict segmentation masks and edges. The segmentation masks and edges are then used to predict finer segmentation masks.

Generative adversarial network-based approaches also have their fair share of contributions in this field. Post-processing techniques such as conditional receptive fields (CRF) are normally used to enhance the results after segmentation. Zhaoa et al. [31] changed this trend by adding CRFs in the end-to-end training with convolution. These convCRF modules are then used in the discriminator part of their generative adversarial network.

In Refs. [32,33], generative adversarial networks (GANs) are used to segment road networks in RS images. GANs contain two main components: a generator and a discriminator. The generator generates segmentation masks to fool the discriminator; on the other hand, the discriminator tries to determine whether the input segmentation mask is from the dataset or was generated by the generator. Shamsolmoali et al. [33] incorporates a spatial pyramid pooling to segment objects at multiple scales. Table 1 portrays a summary of existing models.

Table 1. Summary of existing models.

Reference	DL Method	Main Steps	Findings
Henry et al. [15]	Fully Convolutional Neural Networks (FCNNs)	Segmentation with FCNNs Adjusting the FCNNs for road segmentation.	(1) FCNNs are an effective method to extract roads from SAR images. (2) Adding a tolerance rule to FCNNs can handle mistakes spatially and enhance road extraction.
Xin et al. [16]	DenseUNet	(1) Encoder–Decoder. (2) Backpropagation to Train Multilayer Architectures. (3) DenseUNet.	(1) Combination of dense connection mode and U-Net to solve the problem of tree and shadow occlusion. (2) Use of weighted loss function to emphasize foreground pixels. (3) Dense and skip connection help transfer information and accelerate computation.
Chen et al. [17]	Dense Feature Pyramid Network (DFPN)	(1) Data Preprocessing. (2) DFPN-Based Deep Learning Model. (3) Constructions of Feature Extraction Framework. (4) Establishments of FPN, and DFPN. (5) Generation of Object Proposals. (6) Road Marking Instance Segmentation.	(1) Introduction of the focal loss function in the calculation of mask loss to pay more attention to the hard-classified samples with less-pixel foreground. (2) Combining the "MaskIoU" method into optimizing the segmentation process to improve the accuracy of instance segmentation of road markings.

Table 1. *Cont.*

Reference	DL Method	Main Steps	Findings
Brewer et al. [27]	Transfer learning models: ResNet50, ResNet152V2, Inceptionv3, VGG16, DenseNet201, InceptionResNetV2, and Xception.	(1) Collect data from the cabin of vehicles. (2) Categorize data into groups: high, mid, and low quality. (3) Label data. (4) Classify road segments using transfer learning models. (5) Test the networks on a subset of the Virginia dataset used for training. (6) Test the transfer learning model with Nigerian roads not used in training.	(1) Capture of variance in road quality across multiple geographies exploration of different DL approaches in a wider range of geographic contexts. (2) Need for more tailored approaches for satellite imagery analysis. (3) Fuzzy-class membership for object qualification using satellite data and CNNs. (4) Continuous estimation of road quality from satellite imagery. (5) Use of a phone app combined with ML for road quality prediction.
Heidler et al. [8]	HED-UNet	(1) Computation of pyramid feature maps using Encoder. (2) Combination of feature maps by the task-specific merging heads using the hierarchical attention mechanism.	(1) Uses hierarchical attention maps to merge predictions from multiple scales. (2) Exploitation of the synergies between the two tasks to surpass both edge detection and semantic segmentation baselines.
Shao et al. [26]	CNN termed multitask road-related extraction network (MRENet)	(1) ResBlock operates according to two steps: extraction of image features using convolution operation and enlarging the receptive field using pooling operation. (2) Pyramid scene parsing (PSP) integrates multilevel features. (3) Multitask learning.	(1) Two-task and end-to-end CNN to bridge the extraction of both road surface and road centerlines by enabling feature transferring. (2) Use of a Resblock and a PSP pooling module to expand the receptive field and integrate multilevel features and to acquire much information.

3. Background Knowledge

3.1. Attention Mechanism

Attention Mechanism is a way to mimic the ability of the human brain to focus on a specific region more than the rest of the image [34–37]. It gives more importance to the features of higher significance and fades the ones of lower priority. Attention mechanisms have been used in various ways throughout the field of deep learning. Originally, attention was used with sequential data to encode the information from all time-steps. In natural language processing (NLP) tasks, the self-attention mechanism is famous for establishing a word with every other word in the sentence, creating a sense of how important each word is compared to every other word. This technique has been extremely popular in NLP tasks and has recently been used in computer vision and surpassing traditional convolutional neural networks.

In encoder–decoder structures, attention has previously been used to quantify the amount of attention the decoder will place on each encoded feature.

The authors in [9] pass in the input image in three scales, and the attention for each scale is learned by the network. The attention generated at larger scales focused more on small details, while the attention at smaller scales focused more on more significant structures, enabling the network to segment objects of all sizes.

3.2. Receptive Field and Spatial Resolution

The receptive field is one of the essential concepts in CNNs [38–40]. Receptive fields can be described as the neural network's ability to see an area in the input image. Large

objects in an image need a lot of contextual information for a convolutional neural network to detect or segment them appropriately. In a case where the object occupies most of the image, a small receptive field will deduce fewer features and thus lead to poor accuracy. Increasing the depth of the network can effectively increase the receptive field; however, such huge networks can be impractical, as they require a lot of memory and processing power. A practical way to solve this issue is to use pooling layers to reduce the spatial dimensions of the features, thus increasing the receptive field. Doubling the receptive field of a network will reduce the spatial dimensions of the features by half. A low spatial resolution can cause small objects to be either over-segmented or missed. The authors in [41] show how dilated convolution layers can reduce the loss of spatial information while still managing to gather distant features (increase in spatial resolution). The architecture D2A U-Net [42] uses dilated convolutions in the model's decoder to improve the receptive field and refine the decoding process. Therefore, along with a large receptive field, it is apparent that a high spatial resolution is also essential to improve segmentation accuracy.

3.3. Dilated Convolution

Before CNNs, image processing was mainly based on filters used to extract useful information. As an example, consider the horizontal edge filter (HEF) given in Equation (1):

$$HEF = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix} \tag{1}$$

To calculate the edges using this filter, it is necessary to calculate the convolutional product of the image and the filter. In a CNN, however, the values of these filters are learned through an iterative process. A convolutional layer has several filters, and the channels belonging to each filter (kernel) must be equal to the number of channels in the input image. This allows a different filter to be applied to each channel.

The main goal of dilating the convolutional operation is to increase the receptive field of the network using the same number of parameters, thereby enabling the segmentation of small and fine edges, as stated in [43]. Dilated convolution can be easily explained as applying a convolution operation on an image using a dilated filter. For example, dilating the filter given by Equation (2) will introduce gaps into the filter (as shown by Equation (3)), thus increasing the receptive field from 3×3 to 5×5, while still using the same nine weights.

$$filter = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 2 & 2 \\ 1 & 1 & 1 \end{bmatrix} \tag{2}$$

$$dilatedfilter = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 2 & 0 & 2 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{bmatrix} \tag{3}$$

4. Proposed Method

This section discusses the proposed methodology in detail. Three problems need to be solved before designing a network to segment a road from very HR images. First, the network should have a receptive field that is large enough to gather the required contextual information. Secondly, if we are to predict segmented pixels, along with accurate edges and segmentation masks, a high spatial resolution is needed, as suggested by [22]. The third problem is how to handle the highly imbalanced dataset effectively.

To address the aforementioned issues, the proposed network has been divided into three parts as shown in Figure 1:

1. The encoder encodes the features in full resolution with the help of attention maps.

2. The encoded features are then used to produce segmentation masks.
3. The previously generated segmentation masks, along with the encoded features, are then used to predict the road edges.

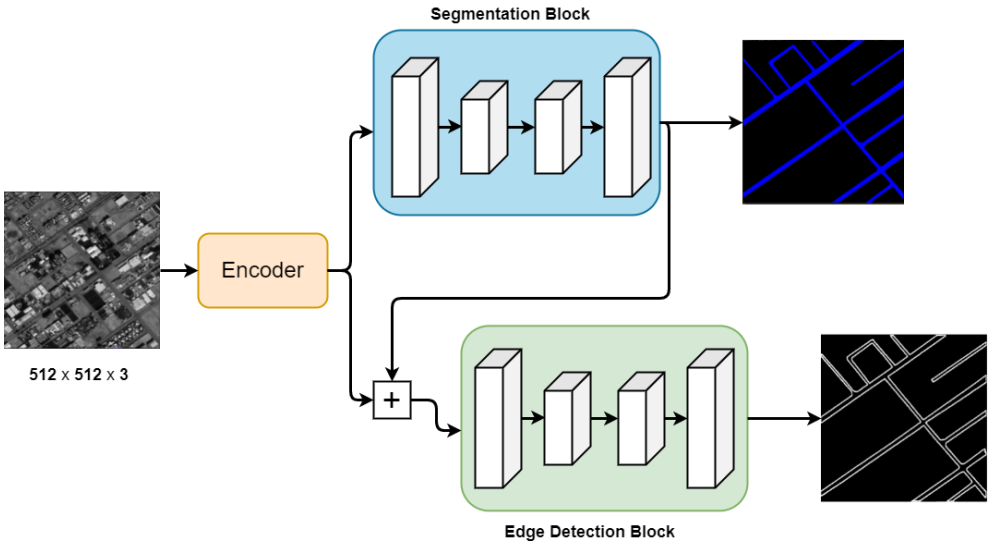

Figure 1. Architecture of the proposed network.

4.1. Encoder

The structure of the encoder shown in Figure 2 contains two spatial and two receptive blocks. The two spatial blocks use dilated convolutional layers to maintain the feature's spatial dimensions while exponentially increasing the receptive field, without increasing the number of trainable parameters. Each spatial block contains three dilated convolutional layers with dilation rates of 1, 2 and 4, respectively. Using dilated convolution with dilation rates in this order increases the receptive field by 15×15, without losing spatial information. This enables the system to predict the segmentation masks in much finer detail.

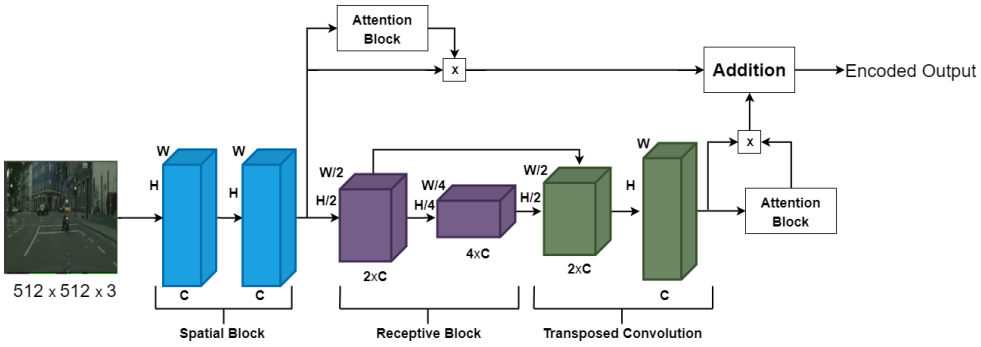

Figure 2. Structure of the proposed encoder.

A large receptive field is required when detecting or segmenting large objects, as more contextual information is required. To further increase our network's receptive field, we decided to use max-pooling layers in our receptive blocks. The first reason why we chose not to completely rely on dilated convolution to increase the receptive field is that, although dilated convolutional layers exponentially increase the theoretical receptive field, the actual receptive field is usually a bit smaller; max-pooling layers, on the other hand, are good at increasing the effective receptive field of the network. The second reason is that going deep with high spatial dimensions can introduce latency. Transposed convolution layers are then used to restore the spatial dimensions of the features from the receptive blocks.

Attention Blocks

When segmenting roads, not all of the pixels in an HR satellite image are equally important; some of the pixels in the image have a higher influence on correct mask prediction than others. This issue is addressed by the attention blocks that generate a 2D attention map equal to the input size. The final convolutional layer in the attention block is followed by a sigmoid activation function shown in Equation (4).

$$Sigmoid(x) = \frac{1}{1 + \exp -x} \tag{4}$$

The sigmoid function maps the input X in the range from 0 to 1, thereby indicating the importance of each pixel in the input. In this case, the attention maps refine the features from the spatial and receptive blocks before adding them to each other. Equations (5)–(7) below show the attention mechanism.

$$sf^* = A(sf) \times sf \tag{5}$$

$$rf^* = A(rf) \times rf \tag{6}$$

$$\text{Encoded output} = sf^* + rf^*, \tag{7}$$

where sf^* and rf^* show the refined features of the spatial and receptive blocks, respectively, and $A(.)$ represents the attention block.

4.2. Segmentation and Edge Detection

The road segmentation network and the edge detection network both operate on full-resolution encoded features and have the structure shown in Figure 3. A skip connection has been used to link the features before max pooling and after the transposed convolution. Skip connections are good at providing the gradient with an alternative way to flow during backpropagation; they are mainly used to counter vanishing gradients. The encoded features of the shape B × C × H × W are fed directly into the segmentation block, where B denotes the batch size, C denotes the number of channels, H denotes the height and W denotes the width. The output from the segmentation block is of shape B × 2 × H × W and the shape of the encoded features in B × 128 × H × W. The encoded features and the segmentation masks are added together; this requires both of them to have the same shape. This problem is addressed by reducing the number of channels of the encoded features by two using a 1 × 1 convolution layer. After adding the predicted segmentation mask and the encoded features with the reduced number of channels, the respective sum is passed through the edge detection block as shown in Algorithm 1. In this situation, detecting edges using only the raw satellite images would have been extremely difficult. However, edge detection gets a lot easier when segmentation information is provided, reducing the need to use huge neural networks to detect edges.

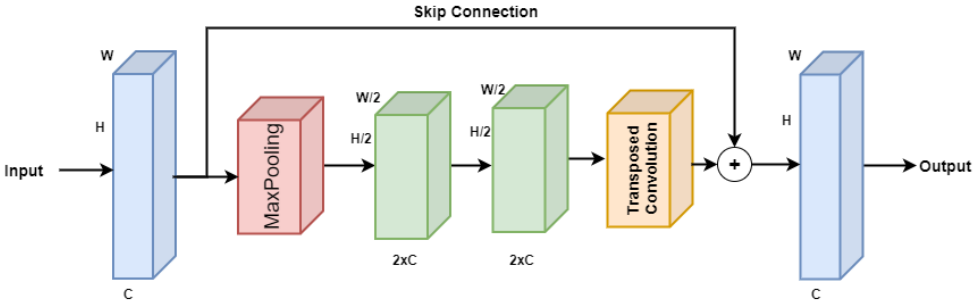

Figure 3. The structure of segmentation and edge detection blocks.

Algorithm 1 Algorithm for road segmentation and edge detection

// x: Input image; rx: Features with reduced channels
// S: Segmented output; E: Edge output
// Comb: Combined channels

Input: Satellite image of size 512 × 512

for x **do**
$x \leftarrow encoder(x)$
$S \leftarrow SegmentationBlock(x)$
// RedConv = 2D convolutional layer with a 1x1 kernel
$rx \leftarrow RedConv(x)$
$Comb \leftarrow S + rx$
$E \leftarrow EdgeBlock(Comb)$
end for

4.3. Loss Function

The ability to properly deal with highly imbalanced data is extremely important when training a neural network, especially in a situation like this (i.e., where only a small area in an image is a road). Here, a combination of a focal Tversky loss and weighted cross-entropy loss has been used as the loss function. The mathematical representation of the Tversky index (TI) is

$$\text{TI} = \frac{TP}{TP + \alpha FN + \beta FP}, \tag{8}$$

where α and β are the weights of false negative and false positive values, respectively. When trained on a highly imbalanced dataset, the network might not predict the road at all yet still have high accuracy. This problem is dealt with by varying the values for α and β; namely, setting α as more significant than β will help the network to penalize false negatives. The focal Tversky loss uses the TI as shown below:

$$\text{Focal Tversky Loss} = (1 - TI)^{\gamma}. \tag{9}$$

Setting γ as more significant than 1 gives a higher loss gradient for samples where the TI is less than 0.5, thereby forcing the model to focus on more complex examples, whereas setting γ as less than one will force the network to focus on those examples where the TI is more significant than 0.5. The overall loss function used in this work is shown in the equation below:

$$\text{Loss function} = \text{Tversky weight} \times \text{Focal tversky loss} + \text{Cross entropy weight} \times \text{Cross entropy loss} \tag{10}$$

5. Materials

Several satellite images representing cities in Saudi Arabia are considered to validate the proposed approach.

5.1. Understanding Saudi Arabia's Land Cover

The Middle East and North Africa's urban centers are undergoing rapid transformation, including significant growth in the youth population and an urbanization trend that will result in large numbers of people settling in major cities. There is a growing recognition that urban spaces in the Middle East, including Saudi Arabia, are in desperate need of more culturally and socially relevant design solutions [44]. Saudi Arabia is a large country divided into 13 administrative provinces that include several major cities, such as Riyadh, Jeddah and Dammam, with varying cultural topography and size of the built environment. In 1937 Saudi Arabia's population was estimated to be less than 5 million people, while

now it is around 31 million people, with a population density of about 15.3 people per square kilometer. Its population growth rate was 2.55% by 2014 [44].

Saudi Arabia is not only a developing country with a scarcity of relevant scientific studies, but it is also a one-of-a-kind case study to investigate. Nearly 88.5 percent of the population is concentrated in urban areas due to industrial, commercial and recreational activities and scarcity of fresh water and agricultural resources. Coastal cities are home to more than half of this segment [44].

A societal change has accelerated the country's transformation into a car-dependent culture; thus, outdoor pedestrians are gradually disappearing from street space. Such phenomena are because of a lack of mass transit systems, the rapid growth of shopping malls, and the failure of urban streets to encourage walking to exist. Moreover, there is currently no information available in Saudi Arabia about pedestrian requirements in the design of modern streets [44].

Saudi Arabia has a high rate of urbanization, so the country faces significant urban challenges today. People have been discouraged from implementing new 'expensive' sustainable housing due to the high cost of living [45].

5.2. Study Regions and Dataset Description

In this study, three regions in Saudi Arabia—namely, Riyadh, Jeddah and Dammam—have been considered, as depicted in Figure 4. The main reason for choosing these three regions is that they are considered the largest cities in Saudi Arabia, which makes the identification of the road network an ideal case to evaluate the effectiveness of the proposed approach.

- Riyadh, Saudi Arabia's capital city, has undergone significant change over the years. Riyadh has been identified as one of Saudi Arabia's fastest-growing cities. The population has increased from 4 million in 2004 to 7 million in 2019. Because of the rate of growth observed in Riyadh, the city is now recognized as one of the fastest-growing in the world by population [46]. Riyadh is located at GPS coordinates of 24°46′27.3540″ N and 46°44′18.9096″ E. Since 1932, the size of what is now known as municipal Riyadh has more than doubled 1000 times, and the population has more than doubled 200 times [46,47]. Aside from traffic issues and pollution, there is a significant social cost associated with the high number of car accidents, which result in one of the world's highest rates of death and casualties. The city's over-reliance on automobiles, combined with a lack of effective street policies aimed at improving walkability, has contributed to a drop in the livability and sustainability scales in Riyadh [47].
- Jeddah, Saudi Arabia's second largest city, has experienced rapid urbanization over the last four decades. Jeddah is located at GPS coordinates of 21°32′35.9988″ N and 39°10′22.0044″ E. Jeddah's population increased rapidly, from nearly 148,000 in 1964 to nearly 3.4 million in 2010, while its urban area increased dramatically, from nearly 18,300 hectares in 1964 to nearly 54,000 hectares in 2007. Furthermore, transportation infrastructure expanded rapidly, from 101 km in 1964 to 826 km in 2007. It has been discovered that approximately 50% of the Jeddah population has limited or no access to the current public transportation system; daily travel behaviors changed, and the daily share of car trips has increased. The proportion of daily car trips increased from 50% in 1970 to 96% in 2012. Jeddah is characterized by deficiency and poor condition of the infrastructure, including buses. High-need districts are concentrated and clustered in the city center, whereas single districts are dispersed to the north of the city center [48].
- Dammam began with a land area of less than 0.7 km^2 in 1947 and grew to 15 km^2 by the 1970s. Its population was estimated to be around 1350 people in 1935 and had grown to 43,000 by 1970, representing a 95.5 percent growth rate during this time period. Between 1950 and 2000, Dammam was one of the world's fastest-growing cities. Dammam has now expanded to over 800 km^2 and has a population of over

1 million people, making it the fifth largest Saudi city in terms of population size. Dammam is situated on a sandy beach with GPS coordinates of 26°26′3.91″ N and 50°06′11.74″ E. Dammam differs greatly from other cities in that it was built almost entirely from the ground up following the discovery of oil. Its urban environment was designed from the beginning with modernist architecture and planning principles in mind, in tandem with rapid advancements in transportation [44].

Experimental results are analyzed using satellite images produced by Spot 7 with a spatial resolution of 1.5 m. Images have been corrected for radiometric and sensor distortions and from acquisition effects. In addition, acquired satellite images have been corrected for viewing angle and ground effects so that they may be superimposed on a map. Moreover, an orthorectification process has been performed to eliminate the perspective effect on the ground.

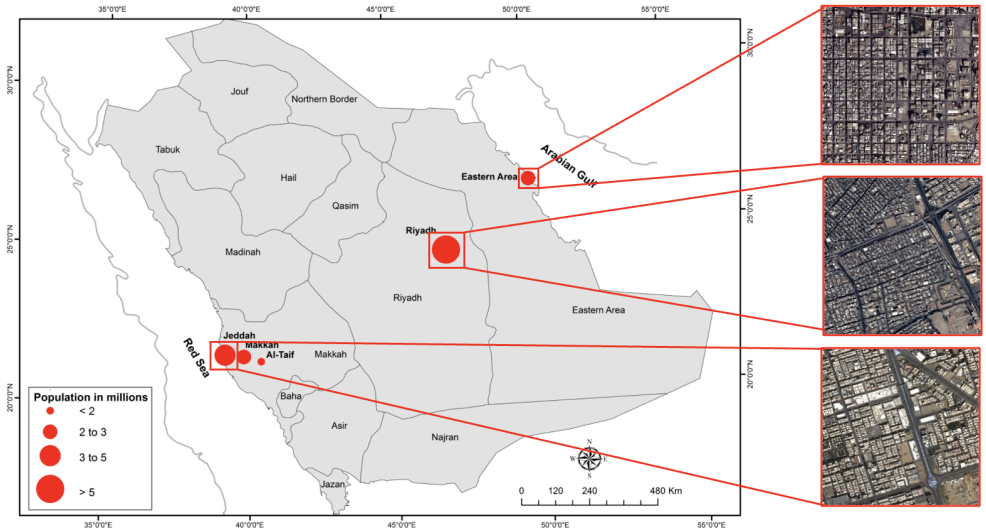

Figure 4. Study regions.

In this paper, experiments are conducted on a dataset composed of 40,488 satellite images having a size of 512 × 512 pixels each.

Figure 5 shows that the dataset is highly imbalanced, and the road covers only a small area.

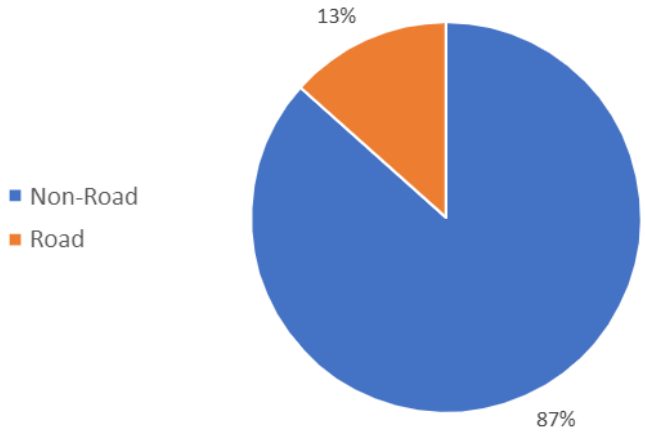

Figure 5. Ratio of road and non-road regions.

5.3. Implementation Details

Practically speaking, it is difficult to train a neural network on an entire satellite image due to the large size of the image. A single satellite image is divided into several

512 × 512 blocks to solve this problem. A small set of the images was manually labeled using VGG Image Annotator. From 900 samples, 800 were for training, 30 for validating, and 70 for testing. The Adam optimizer as well as a learning rate of 1×10^{-5} (0.0067) and a batch size of 5 were used to train the network.

The training process is shown in Figure 6, where the loss function is the combination of weighted cross-entropy loss and the focal Tversky loss. Algorithm 2 shows the training process. The loss between the predicted segmentation mask and the segmentation ground truth is added to the loss between the predicted edge masks and the actual ground truth before backpropagation.

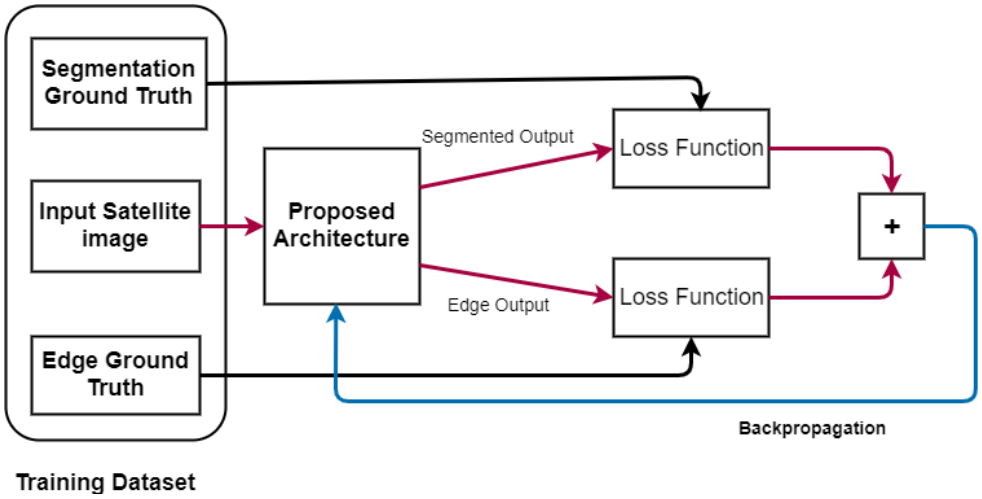

Figure 6. Model training process.

Algorithm 2 Algorithm for forward propagation

// B: Batch size; C: Channel number
// H: Height; W: width
// x: Input sample
// img: Input image
//Sgt: Segmentation groundtruth; Egt: Edge groundtruth
Input:
The entire dataset was first divided into batches in such a way that each sample contained satellite images of shape [B, C, H, W]
Training:

for x **do**
$img, S_{gt}, E_{gt} \leftarrow x$
// Setting the gradients to zero.
optimizer.zero_grad
$encoded \leftarrow Encoder(img)$
$Seg\_predictions \leftarrow Segmentation\_block(encoded)$
$Edge\_predictions \leftarrow Edge\_block(encoded + Seg\_predictions)$
$loss \leftarrow loss\_function(Seg\_predictions, Edge\_predictions, S_{gt}, E_{gt})$
loss.backward()
//updating weights
optimizer.step()
end for

The accuracy of the road edge detection heavily depends on the accuracy of the segmentation. We have used mean intersection over union (mIoU) as the evaluation metric for road segmentation. This metric can help avoid the problem of misleading results provided by pixel accuracy, mainly because the road class representation is small within

the image. Therefore, pixel accuracy measures will be biased in reporting how semantic segmentation will identify the non-road class.

Equation (11) shows how mIoU is calculated, where C is the number of categories, and TP(x), FP(x) and FN(x) are True Positive, False Positive and False Negative values, respectively.

$$\text{mIoU} = \frac{1}{C}\sum_{x=1}^{C}\frac{TP(x)}{TP(x)+FP(x)+FN(x)} \tag{11}$$

6. Results

6.1. Data Augmentation

When training neural networks with a large number of trainable parameters, it is important to have a large dataset to prevent overfitting. Data augmentation is a way to increase the number of samples in a dataset using existing samples. We have rotated our input images and segmentation ground truths at 90 degrees in either direction and flipped the samples on both the x and y axes to increase the size of our dataset by five times. Figure 7 shows an example of data augmentation operations.

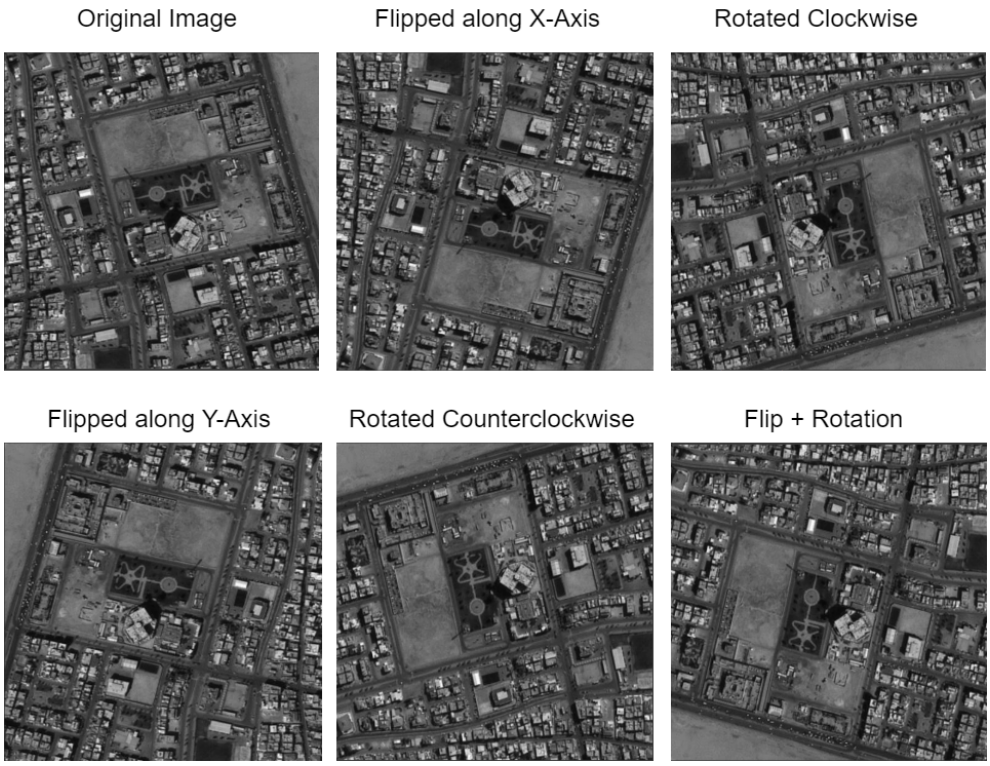

Figure 7. A comparison of the original image with its flipped and rotated counterparts.

6.2. Training the Proposed Model

Weighted cross-entropy loss worked perfectly at the start of the training and is good at dealing with unbalanced data, while adding it to the focal Tversky loss can help at the end, as shown in Figure 8.

Using only cross-entropy loss limited the model segmentation accuracy to 70% mIoU. We added the focal Tversky loss to the weighted cross-entropy loss to overcome it, as explained in the previous section.

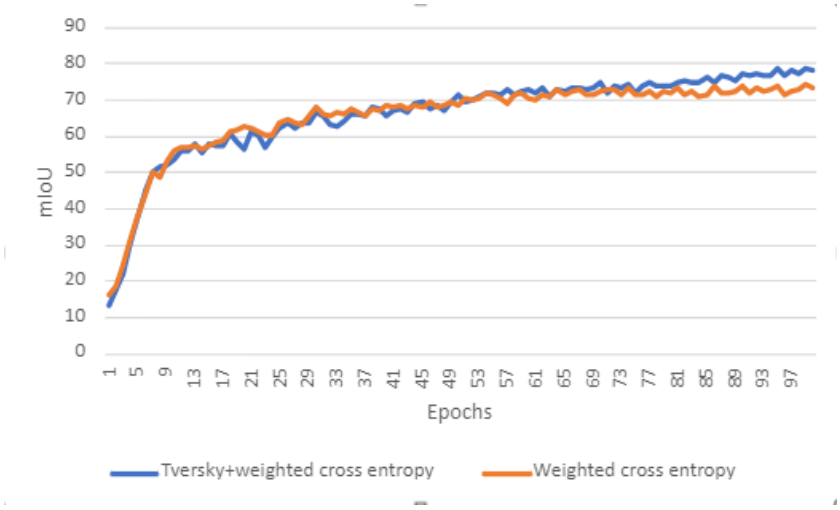

Figure 8. A comparison of weighted cross-entropy and focal Tversky + weighted cross entropy loss.

In Figure 9a, the blue graph shows the linear nature of the focal Tversky loss when *gamma* = 1, while the red graph shows its non-linear nature when *gamma* = 4/3. Figure 9b shows the gradient of the focal Tversky loss function while *gamma* = 4/3. Keeping γ as greater than 1 will give a higher gradient loss when the Tversky index < 0.5, thus forcing the model to focus on harder examples.

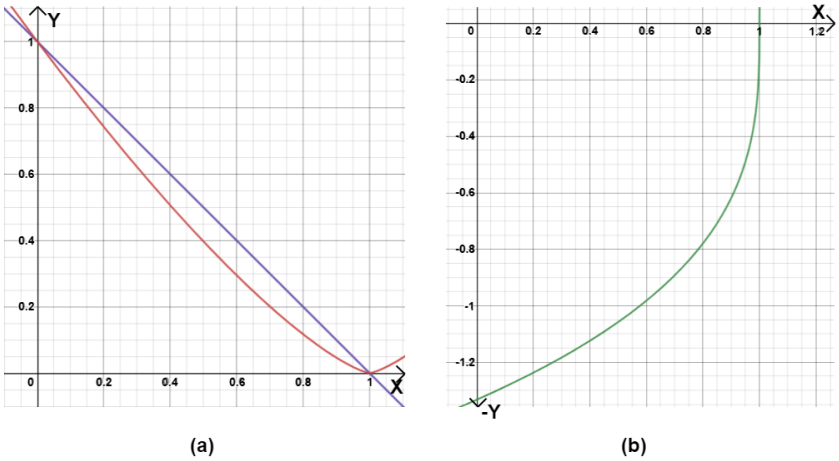

Figure 9. Graph showing the effect of gamma on the gradient of focal Tversky loss. (**a**) The red graph shows the non-linear nature with *gamma* = 4/3. (**b**) The green graph shows the gradient of the red graph when *gamma* = 4/3.

The network has been trained in two stages. The first stage is in the range $0 <$ Epoch ≤ 75, while the second stage ranges from epoch 76 to 100. Tables 2 and 3 show the values used for each of the stages. The experiments indicate that too large or too small values of γ can cause the model to overfit, depending on the size of the dataset. This is the reason that the value of γ is chosen to be close to 1.

6.3. Ablation Study

To study the effectiveness of our attention blocks, we trained our model without the attention blocks and compared its performance against the model trained with the attention blocks; the encoder without the attention blocks is shown in Figure 10. Both the models, in this case, are trained using cross-entropy loss only. The results of both the models were quite similar, with the only difference being a slight 2% improvement in the IOU of the network with the attention blocks. This shows that refining the features using attention

blocks before adding them together can cause a minor improvement in the network's overall performance.

Table 2. Parameters used to train the network when Epoch $\leq$ 75.

Parameters	Value
α	0.7
β	0.3
γ	4/3
Tversky weight	0.2
Cross-entropy weight	0.8

Table 3. Parameters used to train the network when Epoch $>$ 7.5.

Parameters	Value
α	0.5
β	0.5
γ	4/3
Tversky weight	0.5
Cross-entropy weight	0.5

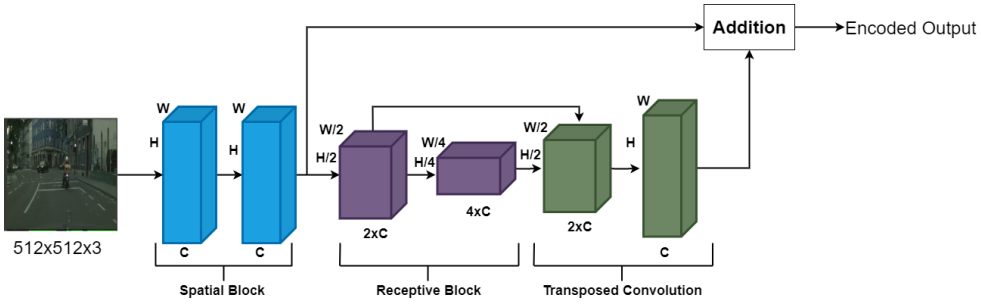

Figure 10. Encoder without the attention blocks.

6.4. Evaluation of the Proposed Approach

Figure 11 shows a comparison of the segmentation performance offered by FCN, SegNet, UNet and our proposed architecture. All four models have been trained using combined weighted cross-entropy and the focal Tversky loss. Figure 12 presents a visual comparison of road segmentation achieved by four models, and Figure 13 shows the road edges detected by our network using the combination of the encoded features and the segmentation masks. It is imperative to note that, when using the combined loss functions, the proposed model does not settle at a 70% mIoU like it previously did in Figure 8.

Table 4 shows that our proposed network has achieved superior performance than SegNet and FCN. The performance of UNet, on the other hand, is almost the same as our proposed technique. UNet has an IOU of 77.34%, while our proposed method has an IOU of 78.3%. It is also worth noting that the size and the learnable parameters of UNet are a lot more than our proposed network. This research goes against the traditional way using back-to-back convolutional block and downsampling the features after each step, and it presents the idea of developing an encoder with separate modules where each module is specialized for each task. Therefore, we can assume that our technique offers a more cost-effective way to achieve reasonably good results for segmenting roads in RS images.

In this paper, we are dealing with detecting roads, which are, in general, small objects with regards to other objects in satellite images such as vegetation, sea and urban areas. Working with higher spatial resolution can better help deep learning methods identify road vs. non-road classes. On the other hand, the multispectral data will improve the results since there is more information provided to deep learning to better identify objects. Working with low spatial resolution has shown a slight decrease in the accuracy of identifying objects

in images, especially in the case of small details and sharp edges. However, increasing the resolution would require a deeper model with a much higher receptive field.

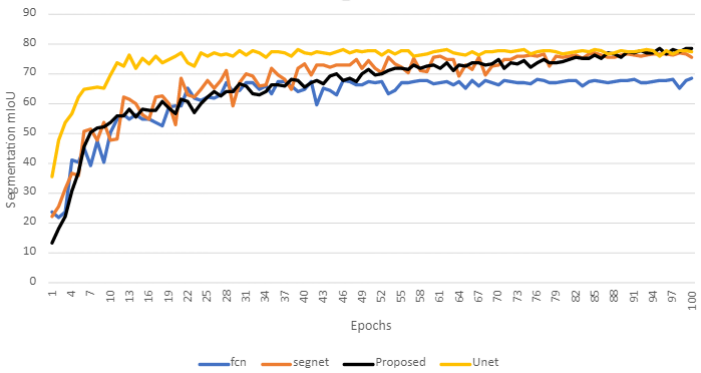

Figure 11. A comparison of the validation accuracy of FCN, SegNet, UNet and our proposed network.

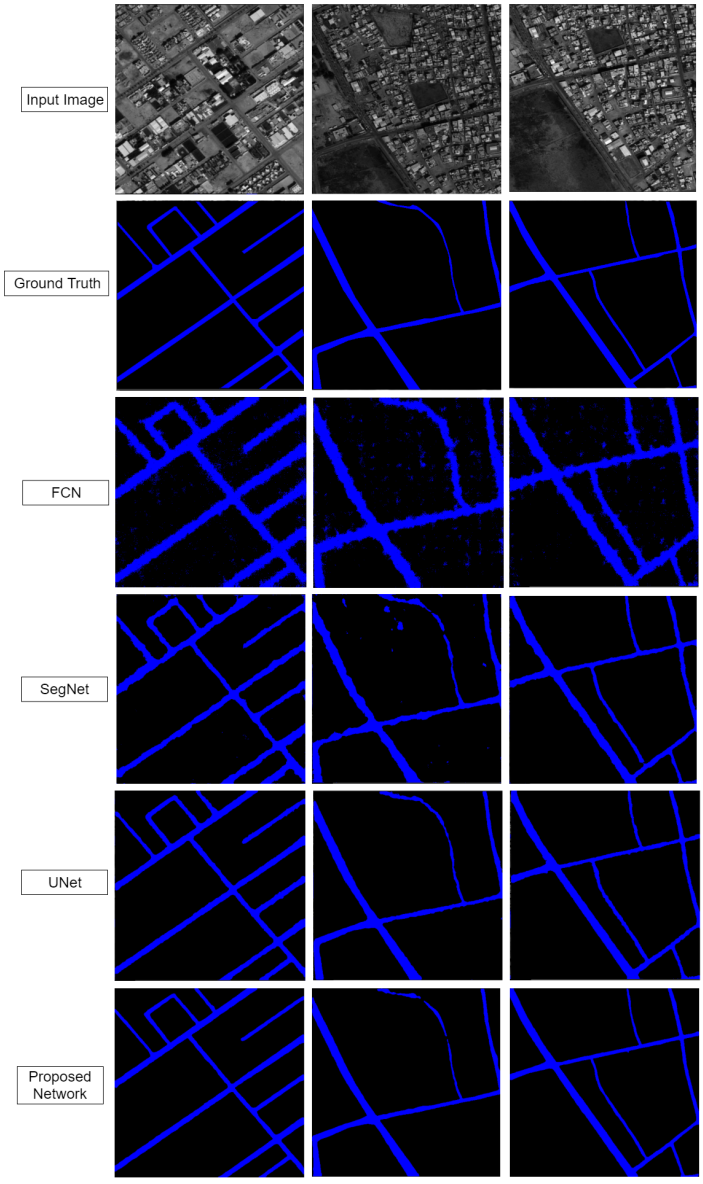

Figure 12. A visual comparison of the segmentation masks offered by FCN, UNet, SegNet and our proposed architecture.

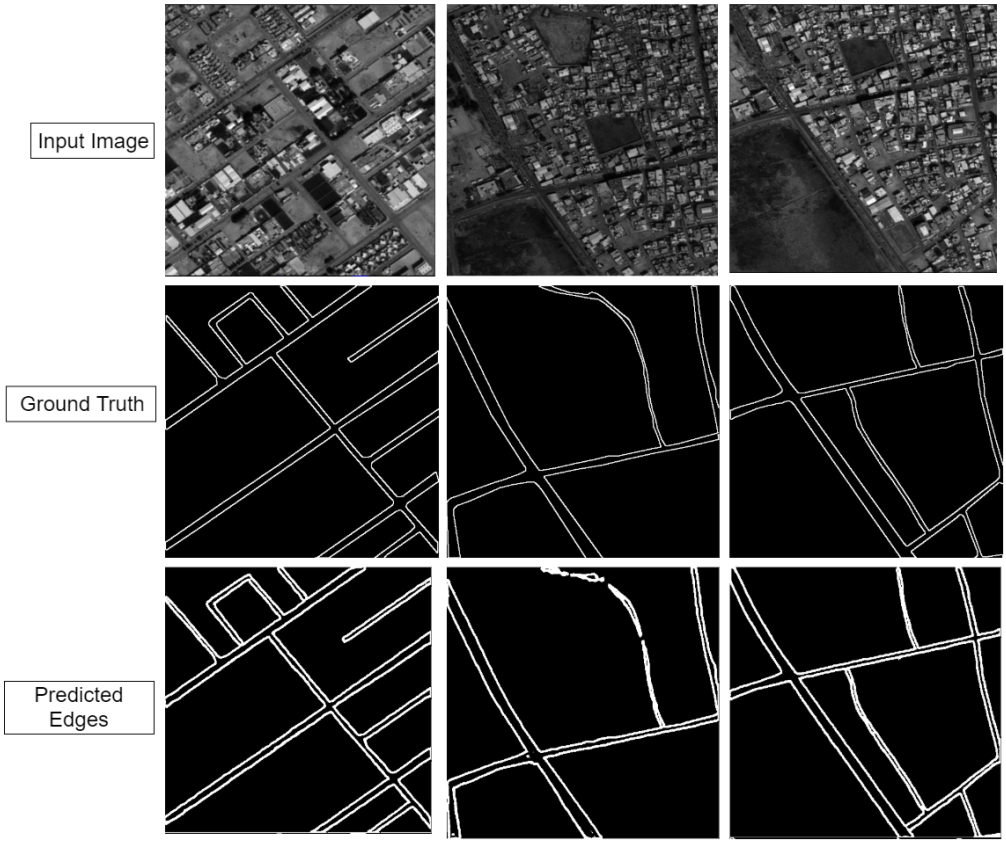

Figure 13. A visual representation of a comparison of the predicted edges and the ground truth.

Table 4. Comparison of the segmentation accuracy offered by FCN, SegNet, UNet and our proposed network.

Network	Size	Learnable Parameters	mIoU	Dice Score
Proposed Network	23 Mb	2.93 M	78.3	87.47
UNet	118.5 Mb	31.04 M	77.34	87.97
SegNet	28.2 Mb	7.37 M	75.44	85.23
FCN	71.1 Mb	18.65 M	68.53	79.83

6.5. Massachusetts Dataset

To better evaluate the performance of the proposed approach, the same experiments conducted for the Saudi Arabia dataset were conducted for the Massachusetts dataset. This dataset contains 1711 images, with 1108 images belonging to the training set, 14 images to the validation set, and 49 images to the test set. Figure 14 shows a visual comparison of the segmentation masks predicted by UNet, SegNet and our proposed architecture. The features jointly encoded by our special blocks and receptive blocks in our encoder have helped our network segment small details. Although the mIoU score of UNet is close to ours, as shown in Table 5, our network is better at segmenting fine details while being much smaller in size, a feature that is important when those masks are being used to predict edges. The dataset was accessed on 7 November 2021 and can be found at the following link: https://www.cs.toronto.edu/~vmnih/data/.

Table 5. Comparison of the segmentation accuracy offered by SegNet, UNet and our proposed network on the Massachusetts dataset.

Network	Size	Learnable Parameters	mIoU	Dice Score
Proposed Network	23 Mb	2.98 M	80.71	89.82
UNet	118.5 Mb	31.04 M	80.32	90.25
SegNet	28.2 Mb	7.37 M	77.74	88.49

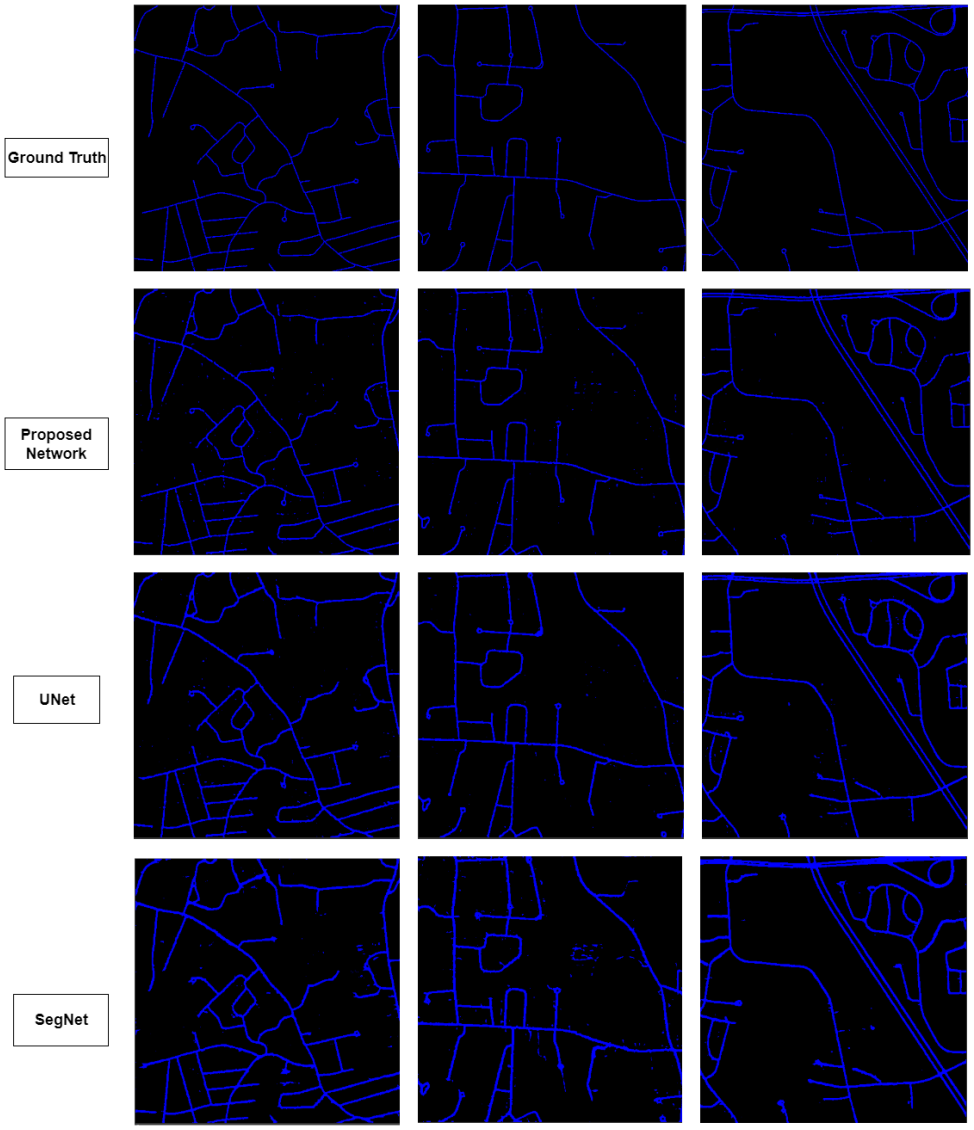

Figure 14. A visual comparison of the segmentation masks predicted by UNet, SegNet and our proposed architecture.

7. Discussion

This paper proposes a technique to encode features to predict fine segmentation masks and road edges using the predicted masks.

High-resolution images can generate enormous amounts of raw data; thus, extracting road information from them is not easy. Manual and traditional approaches to extracting roads are expensive, time-consuming and prone to errors due to human operators, plus the difficulties posed by irregular and complex roads structures remain [13]. On the other hand, automatic road extraction from high-resolution RS images is complicated. Road features such as vehicles, buildings on the roadside and visible tree shadows have similar spectral

values to road pixel values. The lack of context of road parts [49] as well as irregular road segments and complex network structures pose additional challenges [16,50]. In terms of colors, the reflections, patterns and occlusions are similar to road features [13,51].

Furthermore, the road features can be affected by different sensor types, spectral and spatial resolutions, volatile weather conditions, diverse road materials and complex backgrounds. It is critical to analyze the road features and road model in a normal situation without taking into account noise interference [51]. These are the reasons why traditional segmentation techniques such as Unet [3] or FCN [5] are not an ideal solution to the problem. The idea of using high-resolution feature maps is already a well-known solution for predicting fine segmentation masks. This idea is heavily adopted in the technique used by Hamaguchi et al. [22], when authors have designed this technique for the segmentation of small objects such as buildings with low context information to perform the task, entirely relying on dilated convolution to achieve large enough receptive fields. HED-UNET [8], on the other hand, has used a UNet-like architecture with aggressive downsampling using pooling layers to increase the receptive field; this allows their network to have a receptive field as large as the image. In their case, having a large receptive field is extremely important, as segmenting the coastline needs a huge amount of context information. MRENet [26] uses a PSP module after multiple downsampling layers to segment roads of all sizes, but the downsampling, in this case, can damage the rich spatial information. It can be argued that the skip connections used by MRENet to connect layers from the encoder to the decoder are an excellent way to preserve the spatial information from the encoder. However, looking at it differently can show that each convolutional block in the encoder receives the features that the previous convolutional block has already downsampled. This approach with aggressive down and upsampling can result in a bulky model. The same idea can be applied to Segnet [4].

This study proposes a way to deal with these problems; our two-staged encoder plays a considerable role in this issue. The first stage is stacked with multiple dilated convolutional layers encoding the features in full resolution, while the second stage rapidly downsamples the features using pooling layers to increase the receptive field on the network. However, this technique has been shown to solve the problem, but full processing resolution can be memory hungry.

In previous studies, attention maps have been used for various applications. Fu et al. [52] proposed two types of attention modules, position attention module and channel attention module, to capture global and local features. The results from both the attention modules are combined to achieve state-of-the-art segmentation performance. MANet [7] proposes a technique to refine the multi-scale features before combining them. The goal of our attention mechanism was to allow a better addition of features from both parts of the encoder. This enables the network to learn where high-resolution features are helpful and when they need a large receptive field.

Although weighted cross-entropy loss is a great way to deal with the unbalanced dataset, adding it with Tversky loss has been shown to perform well at the end of the training process.

Remote sensing road segmentation labeled data sets are usually small in size, so it is reasonable to consider using semi-supervised learning to use the unlabeled data. Using our feature encoding and edge generation techniques in a generative adversarial network and semi-supervised manner can be an excellent path for future developments. Its applications can also be differentiated between different types of roads.

8. Conclusions

The development of satellite imaging technology has increased the amount of remote sensing data available, which, in turn, has led to a considerable increase in the number of computer vision techniques used to process these images. Methods to process these images have been rapidly evolving, with road segmentation being one of the most critical tasks.

Generating road edges from complex remote sensing data is almost impossible in cases where the shadows from buildings, trees and noise increase the scene's complexity. A straightforward way to solve this problem is to generate segmentation masks and then use the generated segmentation masks to predict the road edges. The effectiveness of this technique heavily depends on the sharpness of the segmentation masks.

This paper introduces a DL-based technique to segment roads and to predict road edges from HR remote sensing images.

Our model uses a hybrid encoder divided into two parts: the first part extracts the features in full resolution, and the second part generates high-resolution feature encoding. In contrast, the second part uses max-pooling layers to increase the overall receptive field of our network, providing the network with enough context information to work with it. A 2D activation map is generated for each part before the features from both parts are added together, allowing the network to select how much attention to give to the features from each stage of the encoder. This facilitates the segmentation of large roads and the generation of fine-edged segmentation masks, which is an essential feature when the performance of the edge detection module depends heavily on the segmentation performance.

The fine-edged segmentation masks and the encoded features are then used to detect the road edges. The segmented masks reveal the structure of a road from a highly complex environment, making it easier to detect the edges.

Good quality data and balanced classes are equally important as the network structure; in our case, the classes are highly imbalanced.

We have experimented with weighted cross-entropy loss and Tversky loss functions and have shown that combining the two can enhance the training performance in a highly imbalanced situation such as this one.

The code is available at the following GitHub link: https://github.com/WadiiBoulila/Semantic-Segmentation-Edge-Detection.

Author Contributions: Conceptualization, W.B. and S.M.; methodology, W.B., S.M. and A.K.; software, S.M.; validation, S.M., W.B. and F.A.; formal analysis, H.G., W.B., S.M., A.K. and J.A.; investigation, H.G.; resources, H.G., W.B., F.A. and J.A.; data curation, W.B.; writing—original draft preparation, H.G., W.B. and S.M.; writing—review and editing, H.G., W.B., A.K., F.A. and J.A.; visualization, W.B. and S.M.; supervision, H.G., W.B., A.K., F.A. and J.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: The author would like to thank Prince Sultan University for their support.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Boulila, W.; Farah, I.R.; Saheb Ettabaa, K.; Solaiman, B.; Ben Ghézala, H. Spatio-Temporal Modeling for Knowledge Discovery in Satellite Image Databases. In Proceedings of the CORIA, Sousse, Tunisia, 18–20 March 2010; pp. 35–49.
2. Boulila, W. A Top-Down Approach for Semantic Segmentation of Big Remote Sensing Images. *Earth Sci. Inform.* **2019**, *12*, 295–306. [CrossRef]
3. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer International Publishing: Munich, Germany, 2015; pp. 234–241.
4. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]
5. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. 2015. pp. 3431–3440. Available online: https://www.computer.org/csdl/proceedings-article/cvpr/2015/07298965/12OmNy49sME (accessed on 19 September 2021). [CrossRef]
6. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:cs.CV/1706.05587.
7. Li, R.; Zheng, S.; Duan, C.; Zhang, C.; Su, J.; Atkinson, P.M. Multi-Attention-Network for Semantic Segmentation of Fine Resolution Remote Sensing Images. *arXiv* **2020**, arXiv:eess.IV/2009.02130.

8. Heidler, K.; Mou, L.; Baumhoer, C.; Dietz, A.; Zhu, X.X. HED-UNet: Combined Segmentation and Edge Detection for Monitoring the Antarctic Coastline. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–14. [CrossRef]
9. Tao, A.; Sapra, K.; Catanzaro, B. Hierarchical Multi-Scale Attention for Semantic Segmentation. *arXiv* **2020**, arXiv:cs.CV/2005.10821.
10. Salehi, S.S.; Erdogmus, D.; Gholipour, A. Tversky Loss Function for Image Segmentation Using 3D Fully Convolutional Deep Networks. In *International Workshop on Machine Learning in Medical Imaging*; Springer: Cham, Germany, 2017; pp. 379–387. [CrossRef]
11. Cira, C.I.; Alcarria, R.; Manso-Callejo, M.Á.; Serradilla, F. A deep learning-based solution for large-scale extraction of the secondary road network from high-resolution aerial orthoimagery. *Appl. Sci.* **2020**, *10*, 7272. [CrossRef]
12. Wan, J.; Xie, Z.; Xu, Y.; Chen, S.; Qiu, Q. DA-RoadNet: A Dual-Attention Network for Road Extraction from High Resolution Satellite Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 6302–6315. [CrossRef]
13. Abdollahi, A.; Pradhan, B.; Shukla, N.; Chakraborty, S.; Alamri, A. Deep Learning Approaches Applied to Remote Sensing Datasets for Road Extraction: A State-Of-The-Art Review. *Remote Sens.* **2020**, *12*, 1444. [CrossRef]
14. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 17–19 June 2017; pp. 6230–6239. [CrossRef]
15. Henry, C.; Azimi, S.M.; Merkle, N. Road segmentation in SAR satellite images with deep fully convolutional neural networks. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 1867–1871. [CrossRef]
16. Xin, J.; Zhang, X.; Zhang, Z.; Fang, W. Road extraction of high-resolution remote sensing images derived from DenseUNet. *Remote Sens.* **2019**, *11*, 2499. [CrossRef]
17. Chen, S.; Zhang, Z.; Zhong, R.; Zhang, L.; Ma, H.; Liu, L. A dense feature pyramid network-based deep learning model for road marking instance segmentation using MLS point clouds. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 784–800. [CrossRef]
18. Emara, T.; Munim, H.E.A.E.; Abbas, H.M. LiteSeg: A Novel Lightweight ConvNet for Semantic Segmentation. In 2019 Digital Image Computing: Techniques and Applications (DICTA). 2019. Available online: https://ieeexplore.ieee.org/abstract/document/8945975 (accessed on 19 September 2021). [CrossRef]
19. Paszke, A.; Chaurasia, A.; Kim, S.; Culurciello, E. ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation. *arXiv* **2016**, arXiv:1606.02147.
20. Aich, S.; van der Kamp, W.; Stavness, I. Semantic Binary Segmentation Using Convolutional Networks without Decoders. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Salt Lake City, UT, USA, 18–23 June 2018.
21. Sovetkin, E.; Achterberg, E.J.; Weber, T.; Pieters, B.E. Encoder–Decoder Semantic Segmentation Models for Electroluminescence Images of Thin-Film Photovoltaic Modules. *IEEE J. Photovolt.* **2021**, *11*, 444–452. [CrossRef]
22. Hamaguchi, R.; Fujita, A.; Nemoto, K.; Imaizumi, T.; Hikosaka, S. Effective Use of Dilated Convolutions for Segmenting Small Object Instances in Remote Sensing Imagery. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 1442–1450. [CrossRef]
23. Boulila, W.; Ghandorh, H.; Khan, M.A.; Ahmed, F.; Ahmad, J. A Novel CNN-LSTM-based Approach to Predict Urban Expansion. *Ecol. Inform.* **2021**, *64*, 101325. [CrossRef]
24. Boulila, W.; Mokhtar, S.; Driss, M.; Al-Sarem, M.; Safaei, M.; Ghaleb, F. RS-DCNN: A Novel Distributed Convolutional Neural Networks based-approach for Big Remote-Sensing Image Classification. *Comput. Electron. Agric.* **2021**, *182*, 106014. [CrossRef]
25. Alkhelaiwi, M.; Boulila, W.; Ahmad, J.; Koubaa, A.; Driss, M. An Efficient Approach Based on Privacy-Preserving Deep Learning for Satellite Image Classification. *Remote Sens.* **2021**, *13*, 2221. [CrossRef]
26. Shao, Z.; Zhou, Z.; Huang, X.; Zhang, Y. MRENet: Simultaneous Extraction of Road Surface and Road Centerline in Complex Urban Scenes from Very High-Resolution Images. *Remote Sens.* **2021**, *13*, 239. [CrossRef]
27. Brewer, E.; Lin, J.; Kemper, P.; Hennin, J.; Runfola, D. Predicting road quality using high resolution satellite imagery: A transfer learning approach. *PLoS ONE* **2021**, *16*, e0253370. [CrossRef]
28. Zhang, J.; Wei, F.; Feng, F.; Wang, C. Spatial–Spectral Feature Refinement for Hyperspectral Image Classification Based on Attention-Dense 3D-2D-CNN. *Sensors* **2020**, *20*, 5191. [CrossRef]
29. He, C.; Li, S.; Xiong, D.; Fang, P.; Liao, M. Remote Sensing Image Semantic Segmentation Based on Edge Information Guidance. *Remote Sens.* **2020**, *12*, 1501 . [CrossRef]
30. Xie, S.; Tu, Z. Holistically-Nested Edge Detection. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1395–1403. [CrossRef]
31. Zhaoa, Z.; Wang, Y.; Liu, K.; Yang, H.; Sun, Q.; Qiao, H. Semantic Segmentation by Improved Generative Adversarial Networks. *arXiv* **2021**, arXiv:2104.09917.
32. Cira, C.I.; Manso-Callejo, M.Á.; Alcarria, R.; Fernández Pareja, T.; Bordel Sánchez, B.; Serradilla, F. Generative Learning for Postprocessing Semantic Segmentation Predictions: A Lightweight Conditional Generative Adversarial Network Based on Pix2pix to Improve the Extraction of Road Surface Areas. *Land* **2021**, *10*, 79. [CrossRef]
33. Shamsolmoali, P.; Zareapoor, M.; Zhou, H.; Wang, R.; Yang, J. Road Segmentation for Remote Sensing Images Using Adversarial Spatial Pyramid Networks. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 4673–4688. [CrossRef]
34. Liu, X.; Milanova, M. Visual attention in deep learning: A review. *Int. Rob. Auto. J.* **2018**, *4*, 154–155.
35. Li, X.; Zhang, W.; Ding, Q. Understanding and improving deep learning-based rolling bearing fault diagnosis with attention mechanism. *Signal Process.* **2019**, *161*, 136–154. [CrossRef]

36. Chen, Y.; Peng, G.; Zhu, Z.; Li, S. A novel deep learning method based on attention mechanism for bearing remaining useful life prediction. *Appl. Soft Comput.* **2020**, *86*, 105919. [CrossRef]
37. Niu, Z.; Zhong, G.; Yu, H. A review on the attention mechanism of deep learning. *Neurocomputing* **2021**, *452*, 48–62. [CrossRef]
38. Luo, W.; Li, Y.; Urtasun, R.; Zemel, R. Understanding the Effective Receptive Field in Deep Convolutional Neural Networks. In Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 4905–4913.
39. Liu, Y.; Yu, J.; Han, Y. Understanding the Effective Receptive Field in Semantic Image Segmentation. *Multimed. Tools Appl.* **2018**, *77*, 22159–22171. [CrossRef]
40. Chen, X.; Li, Z.; Jiang, J.; Han, Z.; Deng, S.; Li, Z.; Fang, T.; Huo, H.; Li, Q.; Liu, M. Adaptive Effective Receptive Field Convolution for Semantic Segmentation of VHR Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 3532–3546. [CrossRef]
41. Liu, R.; Cai, W.; Li, G.; Ning, X.; Jiang, Y. Hybrid Dilated Convolution Guided Feature Filtering and Enhancement Strategy for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
42. Zhao, X.; Zhang, P.; Song, F.; Fan, G.; Sun, Y.; Wang, Y.; Tian, Z.; Zhang, L.; Zhang, G. D2A U-Net: Automatic Segmentation of COVID-19 CT Slices Based on Dual Attention and Hybrid Dilated Convolution. *Comput. Biol. Med.* **2021**, *135*, 104526. [CrossRef] [PubMed]
43. Chen, K.b.; Xuan, Y.; Lin, A.j.; Guo, S.h. Lung Computed Tomography Image Segmentation based on U-Net Network Fused with Dilated Convolution. *Comput. Methods Programs Biomed.* **2021**, *207*, 106170. [CrossRef] [PubMed]
44. Alabdullah, M.M. Reclaiming Urban Streets for Walking in a Hot and Humid Region: The Case of Dammam City, the Kingdom of Saudi Arabia. Ph.D. Thesis, University of Edinburgh, Edinburgh, UK, 2017.
45. Susilawati, C.; Surf, M.A. Challenges facing sustainable housing in Saudi Arabia: A current study showing the level of public awareness. In Proceedings of the 17th Annual Pacific Rim Real Estate Society Conference, Gold Coast, Australia, 16–19 January 2011; pp. 1–12. Available online: http://www.prres.net/ (accessed on 19 September 2021).
46. Alghamdi, A.; Cummings, A.R. Assessing riyadh's urban change utilizing high-resolution imagery. *Land* **2019**, *8*, 193. [CrossRef]
47. Al-Mosaind, M. Applying complete streets concept in Riyadh, Saudi Arabia: Opportunities and challenges. *Urban Plan. Transp. Res.* **2018**, *6*, 129–147. [CrossRef]
48. Aljoufie, M. Spatial analysis of the potential demand for public transport in the city of Jeddah, Saudi Arabia. *WIT Trans. Built Environ.* **2014**, *138*, 113–123.
49. Xu, Y.; Xie, Z.; Feng, Y.; Chen, Z. Road Extraction from High-Resolution Remote Sensing Imagery Using Deep Learning. *Remote Sens.* **2018**, *10*, 1461. [CrossRef]
50. Abdollahi, A.; Pradhan, B.; Alamri, A. VNet: An End-to-End Fully Convolutional Neural Network for Road Extraction From High-Resolution Remote Sensing Data. *IEEE Access* **2020**, *8*, 179424–179436. [CrossRef]
51. Lian, R.; Wang, W.; Mustafa, N.; Huang, L. Road Extraction Methods in High-Resolution Remote Sensing Images: A Comprehensive Review. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 5489–5507. [CrossRef]
52. Fu, J.; Liu, J.; Tian, H.; Fang, Z.; Lu, H. Dual Attention Network for Scene Segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3141–3149.