applied sciences

MDPI

Article

Automated Road Extraction from Satellite Imagery Integrating Dense Depthwise Dilated Separable Spatial Pyramid Pooling with DeepLabV3+

Arpan Mahara *, Md Rezaul Karim Khan, Liangdong Deng, Naphtali Rishe, Wenjia Wang and Seyed Masoud Sadjadi

Knight Foundation School of Computing and Information Sciences (KFSCIS), Florida International University, 11200 SW 8th St CASE 352, Miami, FL 33199, USA; mkhan157@fiu.edu (M.R.K.K.); liadeng@cs.fiu.edu (L.D.); rishen@cs.fiu.edu (N.R.); wwang048@fiu.edu (W.W.); sadjadi@cs.fiu.edu (S.M.S.)
* Correspondence: amaha038@fiu.edu; Tel.: +1-1334-492-0242

Abstract: Road extraction is a sub-domain of remote sensing applications; it is a subject of extensive and ongoing research. The procedure of automatically extracting roads from satellite imagery encounters significant challenges due to the multi-scale and diverse structures of roads; improvement in this field is needed. Convolutional neural networks (CNNs), especially the DeepLab series known for its proficiency in semantic segmentation due to its efficiency in interpreting multi-scale objects' features, address some of these challenges caused by the varying nature of roads. The present work proposes the utilization of DeepLabV3+, the latest version of the DeepLab series, by introducing an innovative Dense Depthwise Dilated Separable Spatial Pyramid Pooling (DenseDDSSPP) module and integrating it in the place of the conventional Atrous Spatial Pyramid Pooling (ASPP) module. This modification enhances the extraction of complex road structures from satellite images. This study hypothesizes that the integration of DenseDDSSPP with a CNN backbone network and a Squeeze-and-Excitation block will generate an efficient dense feature map by focusing on relevant features, leading to more precise and accurate road extraction from remote sensing images. The Results Section presents a comparison of our model's performance against state-of-the-art models, demonstrating better results that highlight the effectiveness and success of the proposed approach.

Keywords: DeepLabV3+; ASPP; deep learning; semantic segmentation; satellite imagery; Xception; remote sensing; Squeeze-and-Excitation; road extraction

Academic Editors: Miguel Angel Patricio and Luis Usero Aragonés

Received: 18 November 2024
Revised: 17 January 2025
Accepted: 19 January 2025
Published: 21 January 2025

Citation: Mahara, A.; Khan, M.R.K.; Deng, L.; Rishe, N.; Wang, W.; Sadjadi, S.M. Automated Road Extraction from Satellite Imagery Integrating Dense Depthwise Dilated Separable Spatial Pyramid Pooling with DeepLabV3+. *Appl. Sci.* **2025**, *15*, 1027. https://doi.org/10.3390/app15031027

1. Introduction

The availability of high-resolution satellite images and development of methodologies to extract roads from satellite images have revolutionized the remote sensing domain. Some sub-domains predominately impacted by this revolution include autonomous navigation, transportation management, urban development planning, and so on. However, the diverse variations in road structures introduce multi-scale characteristics, which in turn lead to limitations in accurate road extraction. Additionally, the sparsity of roads and the presence of shadows in remote sensing images present significant challenges in the automatic and systematic extraction of roads from satellite imagery.

Mnih and Hinton's [1] study presents pioneering work in automatic road detection from high-resolution satellite imagery on a large scale. They effectively incorporated a neural network with millions of trainable weights and successfully equipped the network

to detect road objects in large datasets of urban imagery automatically. Following the work of Mnih, several deep learning methods have been utilized to automate road detection and extraction from remote sensing images. The methods range from general deep learning algorithms [2–4] to task-specific models [5–7].

While these advanced deep learning methods have been successful in semantic segmentation tasks, such as extracting roads from satellite imagery, the downsampling operation in the convolution layers may lead to the loss of spatial information [8], a challenge also noted in feature selection for DDoS detection models [9] and in privacy-preserving federated learning approaches for medical image analysis [10]. To address this problem in semantic segmentation, atrous convolution was utilized in a study by Chen et al. [11], which allows for learning semantic information without spatial loss. This work also proposed an Atrous Spatial Pyramid Pooling (ASPP) module that performed better in capturing multi-scale objects. This approach was expanded in DeepLabV3 [12], which can capture global context, followed by DeepLabV3+ [13], which incorporates a decoder module, with each progression showing improved performance in the semantic segmentation task. Another type of convolution called depthwise separable convolution, if used efficiently, can lead to better performance, as showcased by Chollet et al. [14], who used it in their proposed model called Xception, achieving state-of-the-art performance in a classification task.

Due to its relevance and effectiveness in multi-scale feature capture, the DeepLab series has been explored for road extraction in the recent literature [15–17]. These studies mainly focus on altering the backbone feature extractor in the decoder, altering loss functions, or incorporating new backbone extractors, such as VGG19 [18], ResNet50 [19], and Xception [14], alongside the same ASPP module. Even though the ASPP module has successfully captured multi-scale patterns, it may not adequately capture complex multi-scale features in objects such as roads. The DenseASPP [20], proposed by Yang et al., has effectively captured complex patterns in street scenes due to its capability to interpret dense features. However, despite its effectiveness, DenseASPP tends to be computationally intensive. Considering the contributions and limitations of these studies, our contributions to road extraction are as follows:

- We propose an innovative module called Dense Depthwise Dilated Separable Spatial Pyramid Pooling (DenseDDSSPP) and replace the ASPP module with it in DeepLabV3+.
- We conducted an experimental evaluation of various deep learning models to identify an optimal backbone network, which is Xception.
- The present work integrates the Squeeze-and-Excitation block in the decoder to enable our road extraction process to focus on relevant feature channels from the dense feature map obtained.
- Our study demonstrates the better performance of our proposed model in road extraction compared to state-of-the-art methods across different comparison metrics in a supervised setup.

2. Related Work

Applications of deep learning models have advanced the field of road extraction from satellite imagery. Mnih and Hinton's [1] pioneering work on neural network-based road extraction laid the foundation for subsequent advancements. Their model demonstrated the potential of large neural networks to handle extensive datasets for road detection.

U-Net [21], a well-known model recognized for its efficacy in semantic segmentation, has been widely adapted and extensively applied to road extraction in the recent literature. Zhang et al. [5] introduced Deep Residual U-Net by incorporating residual units into the U-Net architecture, achieving better performance in road extraction compared to its

predecessors, including the original U-Net [21] and Mnih-CNN [1]. Xin et al. [22] proposed a road-extraction deep learning method called DenseUNet, which takes advantage of the U-Net skip connection architecture and dense connections within dense units. DenseUNet is capable of focusing on foreground pixels and achieving comparable results in road extraction by surpassing shadow occlusions to a limited extent. Hou et al. [23] proposed the Complement UNet (C-UNet) model for road extraction from satellite imagery. The method first employs a standard UNet to extract road information, followed by an erasing procedure that removes partially extracted road areas based on a fixed threshold. The erased regions are then processed by Multi-scale Dilated UNet (MD-UNet) to extract finer details, and the outputs from UNet and MD-UNet are fused to produce high-quality road extraction results. Building upon the importance of leveraging prior contextual information through U-Net and spatial structural associations for road extraction, Yang et al. [24] introduced SDUNet, which integrates densely connected blocks with spatial CNN capabilities. The model leverages a structure-preserving module to enhance the extraction of continuous road features by incorporating spatial context in four directions, achieving improved performance in road network extraction.

Continuing the chronological advancements of U-Net, Akhtarmanesh et al. [25] introduced enhancements like a patch-based attention mechanism and rotation-based augmentation to the original U-Net. These innovations led to an advanced-attention U-Net that outperformed earlier methods discussed in this section.

Similarly, LinkNet [26] has been effectively utilized for high-quality road extraction due to its lightweight design and the integration of residual connections. D-LinkNet, proposed by Zhou et al. [27], extended the LinkNet architecture by incorporating a pretrained ResNet34 model [19] as the encoder and adding dilated convolution layers in the central part to enhance the receptive field while preserving spatial details. The decoder remained similar to the original LinkNet, maintaining computational efficiency. One of the recent models for road extraction based on LinkNet is RFE-LinkNet, proposed by Zhao et al. [28]. It incorporates receptive field-enhancement modules to capture long-range dependencies and preserve spatial details. Also, the Channel Attention Module (CAM) and Spatial Attention Module (SAM) have been employed to refine multi-scale features, resulting in improved performance in road extraction.

While the previously mentioned CNN models have demonstrated strong performance in road extraction from satellite imagery, they still face limitations due to the intricate nature of roads. A promising alternative for achieving accurate and high-quality road extraction is the implementation of the DeepLab series. DeepLabV3 [12], for instance, employs atrous convolution within the ASPP module, achieving better results through parallel atrous convolutions adjusted by the output stride, primarily in general semantic segmentation tasks. DeepLabV3+ [13], an extension of DeepLabV3, further enhances segmentation accuracy at object boundaries by incorporating a decoder module. Linghu et al. [17] improved upon DeepLabV3+ by integrating MobileNetV2 as the backbone feature extractor and employing the Dice Loss function, achieving higher overall accuracy in road extraction. However, modifications to the core ASPP module, crucial for multi-scale feature extraction, remain largely unexplored in the context of road extraction. Wu et al. [29] proposed a Dense and Global Spatial Pyramid Pooling (DGSPP) module inspired by the ASPP module but did not take advantage of the encoder–decoder architecture of DeepLabV3+.

Building on these contributions and addressing the limitations of previous studies, this study proposes advancements in road extraction by replacing the ASPP module with DenseDDSSPP in the DeepLabV3+ model [13]. In this approach, the output from a preceding depthwise separable convolution layer, after applying a dilation rate, is merged with the input of the next layer in an iterative procedure. Additionally, our study incorporates

Squeeze-and-Excitation [30] in the decoder module to support the selection of relevant features for road extraction decisions. We hypothesize that this approach will efficiently generate denser features that capture intricate and useful road patterns, potentially leading to improved road extraction from high-resolution satellite imagery.

3. Materials and Methods

This study focuses on extracting roads from high-resolution satellite imagery using an advanced CNN-based model. To achieve this goal, we first revisit the foundational concepts of atrous convolution, depthwise dilated separable convolution, and ASPP. We then detail the proposal and integration of DenseDDSSPP into the DeepLabV3+ architecture. This integration aims to enhance road extraction performance beyond current state-of-the-art models by leveraging the efficiency of dense feature extraction at multiple scales.

3.1. Dilated Convolution in Spatial Pyramid Pooling

Dilated convolution, also called atrous convolution, introduces a dilation rate as a parameter to conventional convolution operations [31], enabling accurate segmentation. This technique allows convolutional filters to expand their receptive field without altering the feature map resolution or increasing the computational cost. Mathematically, the dilated convolution equation is illustrated as

$$Y(f) = \sum_{s=1}^{S} X(f + d \cdot s) \cdot W(s) \tag{1}$$

where $X[f]$ is the input feature map, $W(s)$ is the s-th weight in the filter, S denotes the filter's size, and d denotes the dilation rate. $Y[f]$ is the resultant output feature map. Increasing d enlarges the receptive field, improving the model's pixel-level classification capabilities by capturing broader contextual information. Dilated convolution achieves this improvement by convolving the input X with a filter modified to insert $d - 1$ zeros between consecutive filter values, expanding the filter's coverage area without increasing the number of parameters. This approach prevents the spatial resolution loss commonly associated with the downsampling operations in conventional convolution and helps infer a larger field of view for making road extraction decisions without increasing computation.

3.2. Depthwise Dilated Separable Convolution

Depthwise dilated separable convolution expands the effectiveness of dilated convolution by combining depthwise separability with a dilation rate, enhancing computational efficiency and receptive field coverage (as depicted in Figure 1). This method involves two steps: (a) performing depthwise dilated convolution for each input channel independently with a dilation rate of d, increasing the receptive field to $r_d \times r_d$, and (b) applying pointwise convolution to learn linear combinations of the depthwise convolution outputs [32]. Here, $r \times r$ is the receptive field of a regular convolution for learning representation.

Mathematically, we represent depthwise dilated separable convolution as

$$Y(f,k) = \sum_{c=1}^{C} \left(\sum_{s=1}^{S} X_c(f + d \cdot s) \cdot W_{c,k}(s) \right) \tag{2}$$

where $X_c[f]$ is the input feature map in the channel c, $W_{c,k}(s)$ is the s-th weight in the depthwise filter for the channel c and output channel k, S denotes the filter's size, and d denotes the dilation rate. The inner summation represents the depthwise dilated convolution, and the outer summation represents the pointwise convolution across all input channels, C. $Y(f,k)$ is the resultant output feature map for the k-th output channel.

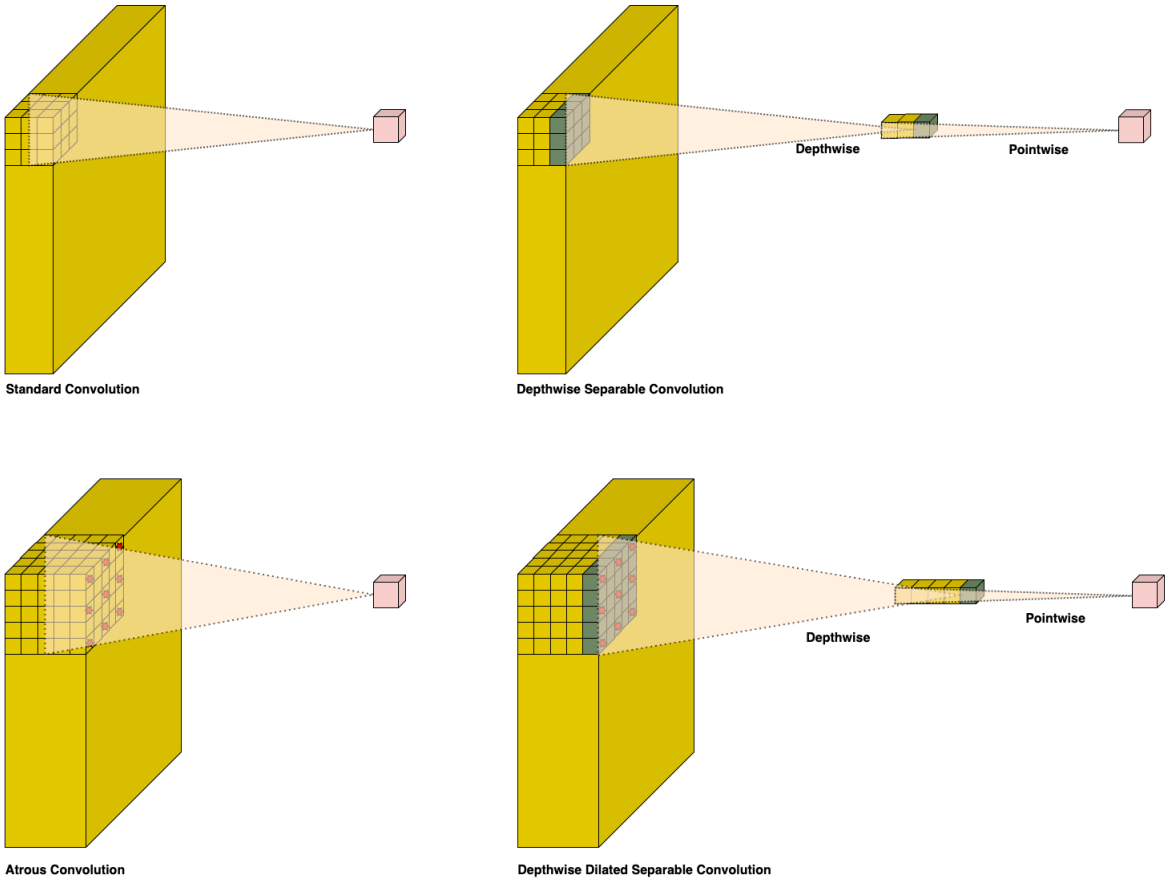

Figure 1. Depiction of different convolutions.

This convolution reduces computational complexity while maintaining an increased field of view for feature representation learning, making it a potentially efficient and powerful tool for road extraction from satellite imagery.

3.3. ASPP Module

The advent of atrous convolution mitigated the issue of spatial resolution loss, but the appearance of variable-sized objects in images introduced new challenges for accurate semantic segmentation. To overcome these challenges, Chen et al. [12] proposed using atrous convolution in a cascading or parallel fashion within the ASPP module, as depicted on the left side of Figure 2, separated by a dotted line. In the cascading mechanism of the ASPP module, the output from a lower atrous layer is passed to a higher layer, producing larger receptive fields. Meanwhile, parallel processing in the ASPP module involves feeding the same input to multiple atrous layers with varying dilation rates. The output obtained from each layer is concatenated to form a comprehensive feature map. This feature map now contains the information of the input across different scales. Mathematically, atrous convolution with $H_{K,d}(x)$ and ASPP can be illustrated as

$$y = H_{3,6}(x) + H_{3,12}(x) + H_{3,18}(x) + H_{3,24}(x) \tag{3}$$

The value of the dilation rates 6, 12, 18, and 24 is based on the output stride [12]. The multi-scale feature aggregation utilizing atrous convolution at various dilation rates in ASPP is a key factor in improving road extraction, particularly at object boundaries.

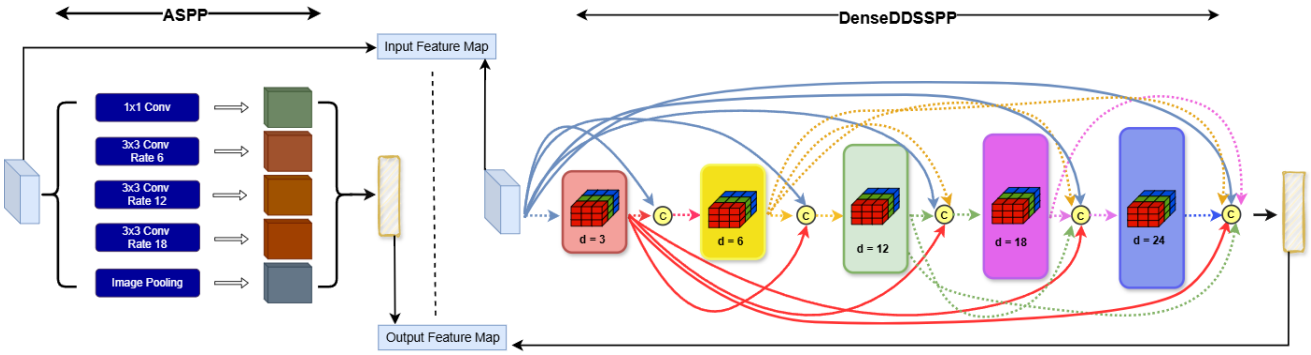

Figure 2. Comparison of architectural designs: ASPP vs. DenseDDSSPP modules.

3.4. DeepLabV3+

DeepLabV3+ [13], an advancement in the DeepLab series, proposes an encoder–decoder architecture. The encoder employs an ASPP module (as discussed in Section 3.3) after processing the input via backbone networks such as VGG19 [18], ResNet50 [19], Xception [14], and so on for a refined feature understanding. Post-ASPP processing, the decoder concatenates the resultant feature with features sourced from the initial stages of the same backbone network. This concatenation can recover object boundaries, leading to better outcomes in semantic segmentation.

3.5. The Proposal and Integration of DenseDDSSPP into the Network

We propose a novel module called DenseDDSSPP and its integration into the DeepLabV3+ architecture, replacing the standard ASPP module in the encoder to achieve more accurate road extraction from satellite imagery (as shown in Figure 3). For a clearer understanding of the differences between ASPP and DenseDDSSPP, Figure 2 provides a side-by-side illustration of the two modules. Motivated by the work of Yang et al. [20], DenseDDSSPP is designed using depthwise dilated separable convolution layers arranged in a cascade, with dilation rates increasing in ascending order. Unlike ASPP, where individual layers operate independently, DenseDDSSPP concatenates the output of each intermediate layer, computed through depthwise dilated separable convolutions with a selected dilation rate, with both the input feature map and the outputs from all previous layers. This dense arrangement facilitates the generation of comprehensive feature maps that integrate multi-scale contextual information. Such an arrangement of depthwise dilated separable convolution layers as a neural network module is a novel contribution in the current literature. The resultant concatenated feature map is sequentially passed to subsequent layers, with this iterative process performed for all convolution layers in the DenseDDSSPP module, resulting in the final dense feature map (depicted in Figure 3, within the encoder section, illustrated with different colored arrows). This approach leverages the computational efficiency of depthwise separable convolutions, as discussed in Section 3.2, to construct DenseDDSSPP as an efficient module. Tracing back to Equations (2) and (3), the layered approach in DenseDDSSPP is mathematically illustrated as follows:

$$Y_l = D_{S,d_l}([Y_{l-1}, Y_{l-2}, \ldots, Y_0]) \tag{4}$$

where l denotes the layer index, d_l is the dilation rate for the layer l, and concatenation is denoted by $[\ldots]$. The expression $[Y_{l-1}, \ldots, Y_0]$ denotes the feature map resulted by merging the outputs from all preceding layers. Here, D_{S,d_l} represents the depthwise dilated separable convolution operation with the kernel size S and dilation rate d_l. These above modifications and integration provide two explicit advantages, i.e., an efficient, denser

feature map and larger receptive field, which are crucial for accurately interpreting complex road structures.

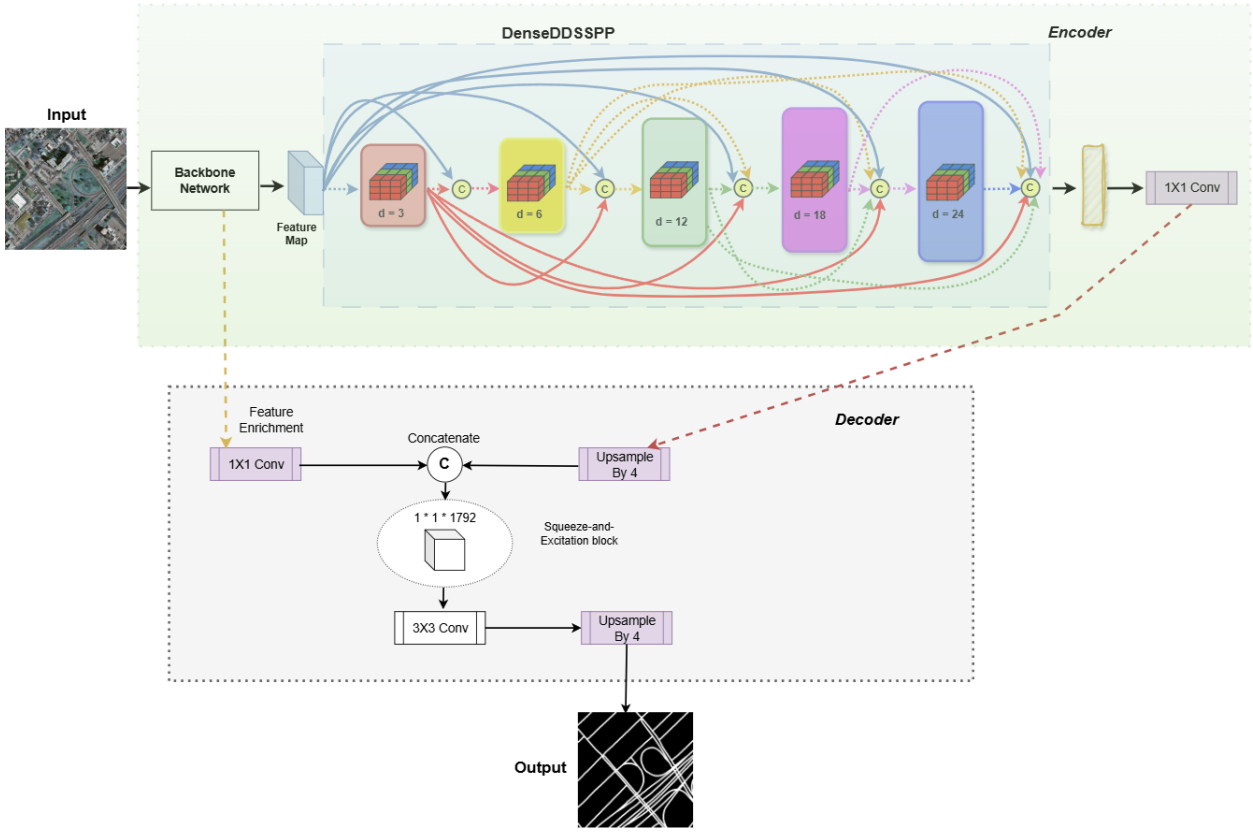

Figure 3. Architecture of DeepLabV3+ with DenseDDSSPP module. The SE block in the decoder operates on a tensor of dimensions 1*1*1792, where 1*1 represents the spatial dimensions and 1792 denotes the number of channels.

3.5.1. Selection of Suitable Backbone Network

As described above, the DenseDDSSPP module requires an input feature map, represented by the light-blue cube in Figure 3, to process through its layers and output a dense feature map. As mentioned in Section 3.4, DeepLabV3+ utilizes a backbone network to generate two feature maps from the input: one, denoted as F_1, serves as the input for the DenseDDSSPP module, and the other, denoted as F_2, is propagated to the decoder (illustrated by the yellow-colored sharp dotted line in Figure 3). Given its important role, selecting a suitable backbone network is essential for the overall performance of the DeepLabV3+ architecture. The present study performed an experimental evaluation of different networks to determine the suitable one. After careful consideration, Xception was selected as the ultimate backbone network for this work. Specifically, the *block*3_*sepconv*2 layer of Xception was chosen to generate F_2, while the *block*13_*sepconv*2_*bn* layer was used to generate F_1. The feature map, F_1, was subsequently fed into the DenseDDSSPP module, producing the enhanced feature map F_1'. A visualization of these layers is presented in Figure 4. Consequently, the process following this step aligns with the latest DeepLab series, i.e., DeepLabV3+ [13], where F_1' and F_2 are concatenated with necessary feature enhancement and upsampling in the decoder to form a multi-scaled and dense feature map, D, as depicted in Figure 3.

Layer (type)	Output Shape	Param #	Connected to
input_3 (InputLayer)	[(None, 512, 512, 3)	0	
block1_conv1 (Conv2D)	(None, 255, 255, 32)	864	input_3[0][0]
.....			
block3_sepconv1_act (Activation	(None, 127, 127, 128	0	add[0][0]
block3_sepconv1 (SeparableConv2	(None, 127, 127, 256	33920	block3_sepconv1_act[0][0]
block3_sepconv1_bn (BatchNormal	(None, 127, 127, 256	1024	block3_sepconv1[0][0]
block3_sepconv2_act (Activation	(None, 127, 127, 256	0	block3_sepconv1_bn[0][0]
block3_sepconv2 (SeparableConv2	(None, 127, 127, 256	67840	block3_sepconv2_act[0][0]
block3_sepconv2_bn (BatchNormal	(None, 127, 127, 256	1024	block3_sepconv2[0][0]
.....			
block13_sepconv1_act (Activatio	(None, 32, 32, 728)	0	add_10[0][0]
block13_sepconv1 (SeparableConv	(None, 32, 32, 728)	536536	block13_sepconv1_act[0][0]
block13_sepconv1_bn (BatchNorma	(None, 32, 32, 728)	2912	block13_sepconv1[0][0]
block13_sepconv2_act (Activatio	(None, 32, 32, 728)	0	block13_sepconv1_bn[0][0]
block13_sepconv2 (SeparableConv	(None, 32, 32, 1024)	752024	block13_sepconv2_act[0][0]
block13_sepconv2_bn (BatchNorma	(None, 32, 32, 1024)	4096	block13_sepconv2[0][0]
conv2d_35 (Conv2D)	(None, 16, 16, 1024)	745472	add_10[0][0]
block13_pool (MaxPooling2D)	(None, 16, 16, 1024)	0	block13_sepconv2_bn[0][0]
batch_normalization_35 (BatchNo	(None, 16, 16, 1024)	4096	conv2d_35[0][0]
add_11 (Add)	(None, 16, 16, 1024)	0	block13_pool[0][0] batch_normalization_35[0][0]
block14_sepconv1 (SeparableConv	(None, 16, 16, 1536)	1582080	add_11[0][0]
block14_sepconv1_bn (BatchNorma	(None, 16, 16, 1536)	6144	block14_sepconv1[0][0]
block14_sepconv1_act (Activatio	(None, 16, 16, 1536)	0	block14_sepconv1_bn[0][0]
block14_sepconv2 (SeparableConv	(None, 16, 16, 2048)	3159552	block14_sepconv1_act[0][0]
block14_sepconv2_bn (BatchNorma	(None, 16, 16, 2048)	8192	block14_sepconv2[0][0]
block14_sepconv2_act (Activatio	(None, 16, 16, 2048)	0	block14_sepconv2_bn[0][0]

Figure 4. A visualization of Xception's layers. The dotted line indicates the omission of intermediate layers for conciseness.

3.5.2. Incorporation of Squeeze-and-Excitation (SE) Block

With the advanced processing mentioned above, we hypothesize that learning the feature map D will enable our model to understand the intricate features of roads and accurately extract high-quality road structures. However, it may encounter limitations in certain edge cases, such as occlusions and the presence of shadows. In these scenarios, an attention mechanism to focus on relevant patterns in edge cases could be invaluable. Previous studies [17,23] have suggested that initial layers primarily capture the rough structure of roads, while deeper layers encode spatial features connected to roads. Building on this understanding, and to leverage both the deeper layers and the DenseDDSSPP module, the resultant feature map D is further enhanced using the Squeeze-and-Excitation (SE) block [30], a technique capable of effectively refining channelwise feature responses (as depicted in Figure 5). This procedure enriches the feature map D to better capture edge cases.

Let D have C channels, where each channel, c, has the height H and width W. The *squeeze* operation is applied, which can be mathematically expressed as

$$z_c = F_{\text{sq}}(d_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} d_c(i, j) \tag{5}$$

This is the global average pooling operation.

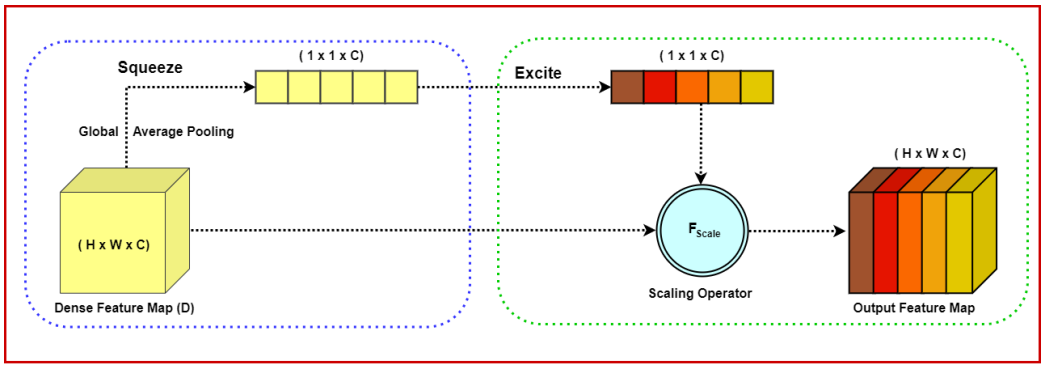

Figure 5. Squeeze-and-Excitation block.

To distinguish between the important channels that lead to better road extraction, we perform the *excitation* operation on the information obtained from the squeeze operation. Similarly to the original work of Hu et al. [30], which acts as an attention mechanism in a channelwise manner, the *excitation* operation can be written as

$$e = F_{\text{ex}}(z, W) = \sigma(W_2 \delta(W_1 z)) \tag{6}$$

where $W_1 \in \mathbb{R}^{C/r \times C}$ is the reduction to the dimension C/r, $W_2 \in \mathbb{R}^{C \times C/r}$ scales back to the dimension C, σ is the sigmoid activation, and δ is the ReLU activation.

Subsequently, we obtain the final feature volume that depicts the importance of one channel over another by a scaling operation. Given the feature map D and the scalar e, the scaling operation can be represented as

$$x_c = F_{\text{scale}}(d_c, s_c) = s_c \cdot d_c \tag{7}$$

where $X = [x_1, x_2, \ldots, x_C]$ and F_{scale} refer to channelwise multiplication.

This completes the enhancement of features through the Squeeze-and-Excitation block in our architecture. Integrating the SE block after the decoder allows the model to capture comprehensive spatial and contextual information while supporting an attention mechanism to focus on relevant patterns, especially in edge cases, resulting in improved road extraction accuracy.

Finally, consistent with the original DeepLabV3+ [13], the above operation is followed by convolution operations and bilinear upsampling, concluding with a sigmoid activation function to ensure the accurate extraction of road structures from satellite imagery.

4. Experimental Results

This section comprises two primary components: the first outlines the datasets and experimental setup, including evaluation metrics and the experimental environment, while the second presents a comparative analysis of the results achieved by our approach against baseline models with the chosen datasets.

4.1. Datasets

In our approach, we used two well-known road datasets, as described below.

4.1.1. Massachusetts Road Dataset

The Massachusetts road dataset, introduced by Mnih and Hinton [1], was selected as the first dataset. This dataset covers roughly 2600 square kilometers of the state of Massachusetts, comprising 1171 satellite images along with their corresponding road-extracted mask images. Each image has dimensions of 1500×1500 pixels, with a resolution of 1 meter per pixel. Given the high dimensions of these images, we employed a patchify procedure, cropping each image to obtain 512×512 tiles. This step was crucial to reduce computational demands and enable the training of complex deep learning models. From the entire dataset, 817 images were selected for our experiments, resulting in 3268 images and corresponding masks of dimensions of 512×512. For training and testing, the dataset was divided using an 80:20 split.

4.1.2. DeepGlobe Road Dataset

Similarly, we used the DeepGlobe dataset [33] as the second dataset to test the performance of our proposed method. It consists of 6226 images with their corresponding road-extracted masks, with a resolution of 1024×1024. To maintain consistency while considering computational demands, we cropped the images and masks into 512×512 tiles, resulting in 24,904 images and 24,904 masks. Due to memory constraints, we selected 9000 images and their corresponding masks and divided the dataset using an 80:20 split into training and testing sets.

4.2. Evaluation Metrics

In this study, we selected commonly used metrics such as Precision, F_1 Score, and Intersection Over Union (IOU) to assess the model's performance. These metrics are defined as follows.

Precision is the ratio of the number of correctly predicted positive observations to the total predicted positive observations:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{8}$$

where TP represents the true positives and FP represents the false positives.

The F_1 Score, which measures the accuracy of positive predictions, is the harmonic mean of Precision and Recall. It can be simplified and expressed mathematically as

$$F_1 = \frac{2TP}{2TP + FP + FN} \tag{9}$$

where FN represents the false negatives.

Intersection Over Union (IOU), which evaluates the pixelwise accuracy of the segmentation, can be mathematically illustrated as

$$\text{IOU} = \frac{TP}{TP + FP + FN} \tag{10}$$

Alternatively, IOU can be understood geometrically as the overlap between the predicted segmentation and the ground truth segmentation.

4.3. Experimental Environment and Baselines

We conducted our experiments using the Python 3.6.8 environment with the TensorFlow framework for all aspects of training and testing. The computational tasks were

carried out using a system comprising eight NVIDIA A100-PCI GPUs, each featuring 80 GB of HBM2 memory. The computation tasks were facilitated using CUDA version 11.8 and NVIDIA driver version 520.61.05. We utilized a mirrored strategy and distributed the computation across four GPU machines to manage the intensive deep learning tasks. All the models were trained for 300 epochs using Adam as the optimizer, with an exponential decay procedure. The initial learning rate was set to 0.001, with the decay maintained at steps of 10,000 and a decay rate of 0.96 to ensure efficient training and better convergence for the models.

Baselines

To validate the efficacy of our proposal, we selected supervised state-of-the-art models in the road-extraction domain. The selected models were U-Net [21], Attention U-Net [25], DeepLabV3+ with ASPP [13], SegNet [34], RFE-Link Net [28], DCSFEP [29], and D-LinkNet [27].

4.4. Comparison and Results

The quantitative results obtained from different state-of-the-art models, including our proposed model, using the Massachusetts road dataset are presented in Table 1, and those from using the DeepGlobe road dataset are presented in Table 2. Models such as U-Net [21], Attention U-Net [25], DeepLabV3+ with ASPP [13], and SegNet [34] were experimented with within our setup, and the metrics' results were reported following training and testing. The results for RFE-Link Net [28] and D-LinkNet [27] and DCSFEP [29] are reported from the original papers due to the relevancy in the setup and usage of the same dataset. As seen in Table 1, our proposed model achieved better performance using the Massachusetts dataset based on the IOU and Precision metrics, with scores of 67.21 and 81.38, respectively, compared to all other models. Additionally, our model showcased better performance compared to all models for the F_1 Score, with 79.29, while the advanced model RFE-Link Net slightly outperformed our model with a score of 80.07.

Similarly, as seen in Table 2, our proposed model was also successful in achieving better performance with the DeepGlobe dataset for the IOU and Precision metrics, with scores of 71.61 and 83.19, respectively, compared to all other models. Along this, the experimental results followed the same trend as obtained in above, in which our model again depicted better performance compared to all models for the F_1 Score, with 81.75, while the advanced model RFE-Link Net slightly outperformed our model with a score of 82.85.

Table 1. Quantitative observation of results obtained from all models in Massachusetts road dataset. Bold values indicate the best performance in the respective metric.

Model	IOU (%)	Precision (%)	F_1 Score (%)
U-Net [21]	64.19	80.23	74.78
Original DeepLabV3+ [13]	65.92	80.04	75.60
SegNet [34]	58.67	78.56	73.73
DCSFEP [29]	62.48	-	76.59
D-LinkNet[27]	63.74	75.89	77.86
RFE-LinkNet [28]	66.77	80.88	**80.07**
Proposed Model	**67.21**	**81.38**	79.29

Table 2. Quantitative observation of results obtained from all models in DeepGlobe road dataset. Bold values indicate the best performance in the respective metric.

Model	IOU (%)	Precision (%)	F_1 Score (%)
U-Net [21]	62.82	80.36	71.83
DeepLabV3+ [13]	69.05	83.16	77.96
SegNet [34]	57.66	80.67	72.09
Attention-UNet [25]	67.42	79.95	80.54
D-LinkNet [27]	63.94	78.54	0.7659
RFE-Link Net [28]	70.72	83.09	**82.85**
Proposed Model	**71.61**	**83.19**	81.75

5. Discussion

In addition to the quantitative metrics' evaluation, our study also presents visual road-extraction depictions from our model, including comparisons to other models. Figure 6 presents the road extraction performed using the validation set of the Massachusetts dataset. As depicted in the diagram, our proposed model achieves accurate road extraction in several parts of the images, as highlighted with red-colored squares, compared to most of the other models.

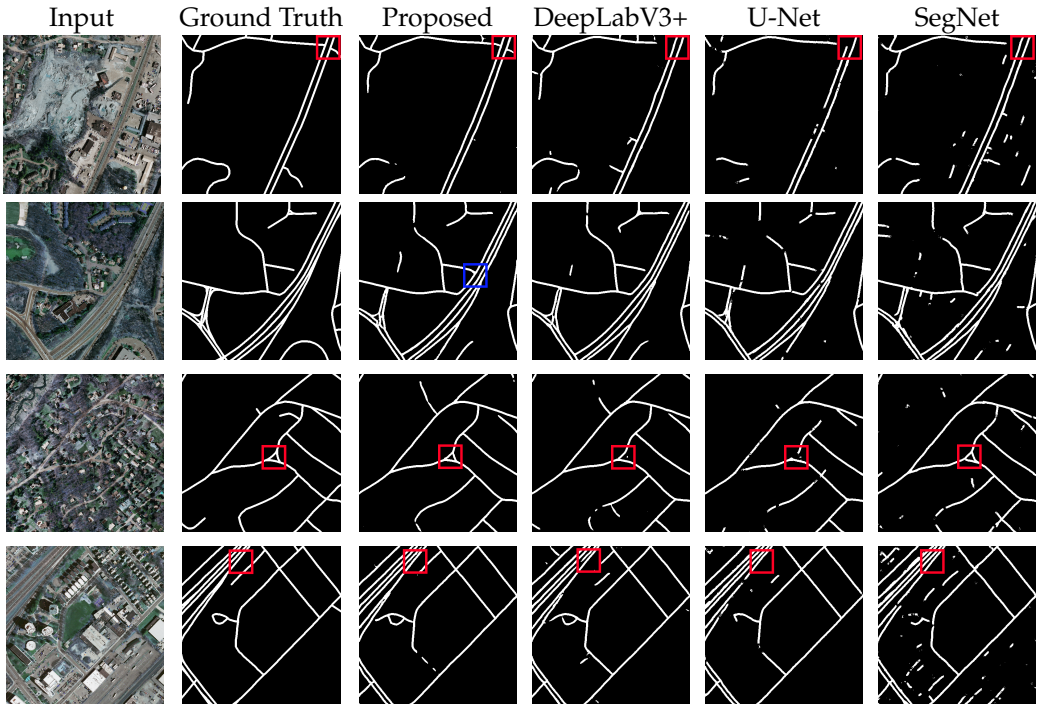

Figure 6. Comparative results of road extraction from the Massachusetts dataset. The figure presents a side-by-side comparison of road extracted by various models, including the proposed model, against the ground truth, highlighting the effectiveness of each approach in synthesizing accurate road extraction.

Similarly, our model demonstrated comparable performance using the DeepGlobe road dataset, as shown in Figure 7. Accurate road construction and connection can be seen in the red-colored squares in the diagram. An interesting observation is that our model tends to connect roads when there are trees present, as highlighted with blue-colored squares in the second row of Figure 6 and the fourth row of Figure 7. In the ground truth images, these road connections are not visible. Upon closer inspection, it can be assumed

that there is a road occluded by trees, and our model may be able to achieve correct road extraction even in such occluded scenarios.

Based on the quantitative and visualization results, our model outperformed various state-of-the-art models in road extraction. Our study not only demonstrated better performance but also presented an efficient computational approach by adopting the depthwise separable mechanism and increasing the field of view with the dilation mechanism. To provide a more intuitive and mathematical understanding of how depthwise separable dilated convolution used in our work brought efficiency compared to standard convolution, we can express the computational savings mathematically.

Given an image of the size $512 \times 512 \times 3$ and using a 3×3 kernel, the standard convolution operation can be expressed as

$$\text{Operations}_{\text{standard}} = H \times W \times K \times K \times C_{\text{in}} \times C_{\text{out}}$$

where $H = 512$ (the height of the image), $W = 512$ (the width of the image), $K = 3$ (the size of the kernel), $C_{\text{in}} = 3$ (the number of input channels), and $C_{\text{out}} = 64$ (the number of output channels). Substituting these values, we obtain

$$\text{Operations}_{\text{standard}} = 512 \times 512 \times 3 \times 3 \times 3 \times 64 = 151,165,440$$

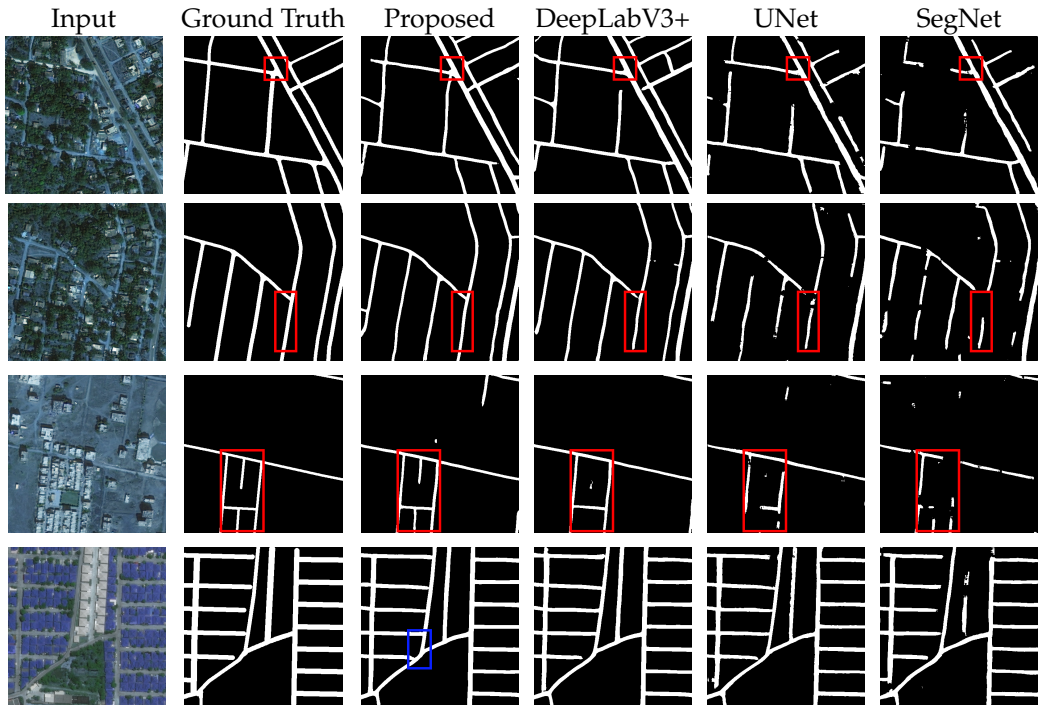

Figure 7. Comparative results of road extraction from the DeepGlobe road dataset. The figure presents a side-by-side comparison of road extracted by various models, including the proposed model, against the ground truth.

Since depthwise separable convolution involves two steps, the operations can be split into two parts, depthwise convolution and pointwise convolution, and their operations are expressed below as

$$\text{Operations}_{\text{depthwise}} = H \times W \times K \times K \times C_{\text{in}}$$

$$\text{Operations}_{\text{pointwise}} = H \times W \times C_{\text{in}} \times C_{\text{out}}$$

Combining both depthwise and pointwise operations, and substituting all the values, the total number of operations for depthwise separable convolution is

$$\text{Operations}_{\text{depthwise separable}} = 7,077,888 + 50,331,648 = 57,409,536$$

To compare the efficiency, we can take the ratio of the operations required:

$$\frac{\text{Operations}_{\text{depthwise separable}}}{\text{Operations}_{\text{standard}}} = \frac{57,409,536}{151,165,440} \approx 0.38$$

This indicates that depthwise separable convolution is approximately 62% more efficient than standard convolution in terms of operations. While existing studies have observed that single-thread depthwise separable convolution often underperforms in accuracy compared to conventional convolution [35], our findings demonstrate that incorporating dense connections can enhance its performance, achieving better results.

Moreover, the computational and memory efficiency of our model is evident when comparing parameter counts and FLOPs with the original DeepLabV3+ [13]. Specifically, our model has a parameter count of 20,874,553 (20.87M), lower than DeepLabV3+ with ASPP, which has 28,314,137 (28.31M) parameters. In terms of FLOPs, our proposed model achieves 160,801,044,036 (160.80G), compared to the 181,132,742,326 (181.13G) required by DeepLabV3+ with ASPP. The reduction in FLOPs reflects a decrease in computational cost, while the smaller parameter count indicates improved memory efficiency.

In summary, the utilization of depthwise separable dilated convolution in our DenseDDSSPP module offers both computational and memory efficiency while also improving feature extraction capabilities. These advancements contribute to the better performance of our proposed model in road extraction tasks. The code implementation for the proposed model, along with the evaluation set for visualization and comparison, is available at https://github.com/amaha7984/Road-Extraction-with-Advanced-Deep-Learning-Model (accessed on 15 January 2025).

6. Conclusions

Road extraction is one of the most important research areas for applications such as autonomous navigation and smart city planning, yet it faces several challenges. Our study addresses these challenges by advancing the capabilities of the DeepLabV3+ model through the introduction of the Dense Depthwise Dilated Separable Spatial Pyramid Pooling (DenseDDSSPP) module, replacing the standard ASPP module. The study also identifies Xception as the optimal backbone network for enhanced feature extraction and integrates the Squeeze-and-Excitation block into the decoder, facilitating channelwise learning to emphasize relevant features. To the best of our knowledge, our work is the first attempt to incorporate Dense Depthwise Dilated Separable Convolution to form a DenseDDSSPP module within a semantic segmentation model for road extraction from satellite imagery. The proposed approach demonstrated better performance using publicly available datasets including the Massachusetts road dataset and the DeepGlobe road dataset. Our model outperformed several state-of-the-art models based on various evaluation metrics in a supervised learning setup, improving the IOU, Precision, and F_1 Score metrics. The visual comparisons further highlighted our model's ability to accurately extract and connect road segments, even in scenarios where roads were occluded by trees.

Future work will focus on enhancing the generalizability of our model by collecting additional datasets and evaluating its performance in zero-shot scenarios [36], enabling the semantic segmentation of road types not encountered during training. We also aim to transition from a supervised setup, which requires annotated images for training, to a

self-supervised learning approach. Inspired by recent advancements in self-supervised learning, such as those by Hou et al. [23] and Mahara et al. [37], we plan to develop methods for road extraction that do not rely on corresponding annotated images for each satellite image. This approach holds the potential to significantly broaden the applicability of our model to diverse and unannotated datasets. Furthermore, while the present work integrates the SE block as the attention mechanism, future work will explore other attention modules, such as CBAM [38] and ECA-Net [39], to perform a comparative analysis and identify a lightweight, optimal attention module for road extraction.

Author Contributions: Conceptualization, A.M.; methodology, A.M.; validation, N.R., L.D., and S.M.S.; formal analysis, A.M., M.R.K.K., and L.D.; investigation, A.M., M.R.K.K., and W.W.; writing—original draft, A.M.; writing—review and editing, N.R., S.M.S., L.D., M.R.K.K., and W.W.; visualization, A.M. and M.R.K.K.; supervision, N.R., L.D., and S.M.S.; funding acquisition, N.R. All authors have read and agreed to the published version of the manuscript.

Funding: This material is based in part upon work supported by the National Science Foundation under Grant Nos. CNS-2018611 and CNS-1920182 and by the Florida Department of Environmental Protection, Grant C-2104.

Data Availability Statement: The datasets used in this article are publicly available datasets for research purposes; the Massachusetts road dataset can be obtained from [1] and the DeepGlobe road dataset can be obtained from [33]. A few examples of images with road extraction in this study are presented in Figures 6 and 7. The complete dataset with road extraction in this research is available upon request from the corresponding author.

Acknowledgments: The authors express their gratitude to the reviewers and editors for their insightful comments and constructive suggestions, which improved the quality and clarity of this paper.

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

References

1. Mnih, V.; Hinton, G.E. Learning to detect roads in high-resolution aerial images. In Proceedings of the Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, 5–11 September 2010; pp. 210–223.
2. Längkvist, M.; Kiselev, A.; Alirezaie, M.; Loutfi, A. Classification and segmentation of satellite orthoimagery using convolutional neural networks. *Remote Sens.* **2016**, *8*, 329. [CrossRef]
3. Sherrah, J. Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery. *arXiv* **2016**, arXiv:1606.02585.
4. Mahara, A.; Rishe, N.D.; Deng, L. The Dawn of KAN in Image-to-Image (I2I) Translation: Integrating Kolmogorov-Arnold Networks with GANs for Unpaired I2I Translation. *arXiv* **2024**, arXiv:2408.08216.
5. Zhang , Z.; Liu, Q.; Wang, Y. Road Extraction by Deep Residual U-Net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753. [CrossRef]
6. Mahara, A.; Rishe, N. Multispectral Band-Aware Generation of Satellite Images across Domains Using Generative Adversarial Networks and Contrastive Learning. *Remote Sens.* **2024**, *16*, 1154. [CrossRef]
7. Gao, X.; Sun, X.; Zhang, Y.; Yan, M.; Xu, G.; Sun, H.; Jiao, J.; Fu, K. An end-to-end neural network for road extraction from remote sensing imagery by multiple feature pyramid network. *IEEE Access* **2018**, *6*, 39401–39414. [CrossRef]
8. Xu, G.; Liao, W.; Zhang, X.; Li, C.; He, X.; Wu, X. Haar wavelet downsampling: A simple but effective downsampling module for semantic segmentation. *Pattern Recognit.* **2023**, *143*, 109819. [CrossRef]
9. Wang, W.; Sadjadi, S.M.; Rishe, N. Curse of Feature Selection: A Comparison Experiment of DDoS Detection Using Classification Techniques. In Proceedings of the 2022 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom), Melbourne, Australia, 17–19 December 2022; pp. 262–269.
10. Das, B.C.; Amini, M.H.; Wu, Y. Privacy risks analysis and mitigation in federated learning for medical images. In Proceedings of the 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Istanbul, Turkiye, 5–8 December 2023; pp. 1870–1873.
11. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef]

12. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
13. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
14. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
15. Quan, B.; Liu, B.; Fu, D.; Chen, H.; Liu, X. Improved deeplabv3 for better road segmentation in remote sensing images. In Proceedings of the 2021 International Conference on Computer Engineering and Artificial Intelligence (ICCEAI), Shanghai, China, 27–29 August 2021; pp. 331–334.
16. Wang, H.; Yu, F.; Xie, J.; Zheng, H. Road extraction based on improved DeepLabv3 plus in remote sensing image. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2022**, *48*, 67–72. [CrossRef]
17. Linghu, Z.; Xiping, Y.; Shu, G.; Lin, H.; Mingyu, Q. An information extraction model of roads from high-resolution remote sensing images based on improved Deeplabv3+. *Remote Sens. Nat. Resour.* **2023**, *35*, 107–114.
18. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
19. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
20. Yang, M.; Yu, K.; Zhang, C.; Li, Z.; Yang, K. Denseaspp for semantic segmentation in street scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3684–3692.
21. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; pp. 234–241.
22. 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]
23. Hou, Y.; Liu, Z.; Zhang, T.; Li, Y. C-UNet: Complement UNet for remote sensing road extraction. *Sensors* **2021**, *21*, 2153. [CrossRef]
24. Yang, M.; Yuan, Y.; Liu, G. SDUNet: Road extraction via spatial enhanced and densely connected UNet. *Pattern Recognit.* **2022**, *126*, 108549. [CrossRef]
25. Akhtarmanesh, A.; Abbasi-Moghadam, D.; Sharifi, A.; Yadkouri, M.H.; Tariq, A.; Lu, L. Road extraction from satellite images using Attention-Assisted UNet. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *17*, 1126–1136. [CrossRef]
26. Chaurasia, A.; Culurciello, E. Linknet: Exploiting encoder representations for efficient semantic segmentation. In Proceedings of the 2017 IEEE Visual Communications and Image Processing (VCIP), St. Petersburg, FL, USA, 10–13 December 2017; pp. 1–4.
27. Zhou, L.; Zhang, C.; Wu, M. D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 182–186.
28. Zhao, H.; Zhang, H.; Zheng, X. RFE-LinkNet: LinkNet with Receptive Field Enhancement for Road Extraction from High Spatial Resolution Imagery. *IEEE Access* **2023**, *11*, 106412–106422. [CrossRef]
29. Wu, Q.; Luo, F.; Wu, P.; Wang, B.; Yang, H.; Wu, Y. 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.* **2020**, *14*, 3–17. [CrossRef]
30. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
31. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv* **2014**, arXiv:1412.7062.
32. Mehta, S.; Rastegari, M.; Shapiro, L.; Hajishirzi, H. Espnetv2: A light-weight, power efficient, and general purpose convolutional neural network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9190–9200.
33. Demir, I.; Koperski, K.; Lindenbaum, D.; Pang, G.; Huang, J.; Basu, S.; Hughes, F.; Tuia, D.; Raskar, R. Deepglobe 2018: A challenge to parse the earth through satellite images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 172–181.
34. 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]
35. Jang, J.G.; Quan, C.; Lee, H.D.; Kang, U. Falcon: Lightweight and accurate convolution based on depthwise separable convolution. *Knowl. Inf. Syst.* **2023**, *65*, 2225–2249. [CrossRef]

36. Bucher, M.; Vu, T.H.; Cord, M.; Pérez, P. Zero-shot semantic segmentation. In *Advances in Neural Information Processing Systems*; Conference Proceedings; Neural Information Processing Systems Foundation, Inc.: La Jolla, CA, USA, 2019; Volume 32, ISBN 9781713807933.
37. Mahara, A.; Rishe, N.D. Generative Adversarial Model Equipped with Contrastive Learning in Map Synthesis. In Proceedings of the 2024 6th International Conference on Image Processing and Machine Vision, Macau, China, 12–14 January 2024; pp. 107–114.
38. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
39. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–13 June 2020; pp. 11534–11542.