



## Article

# MixerNet-SAGA A Novel Deep Learning Architecture for Superior Road Extraction in High-Resolution Remote Sensing Imagery

Wei Wu <sup>1</sup>, Chao Ren <sup>2,\*</sup> , Anchao Yin <sup>2</sup>  and Xudong Zhang <sup>2</sup><sup>1</sup> Power China Guiyang Engineering Corporation Ltd., Guiyang 550081, China; dafoaiyatou@126.com<sup>2</sup> College of Geomatics and Geoinformation, Guilin University of Technology, Guilin 541006, China; yinanchao@glut.edu.cn (A.Y.); 1020211828@glut.edu.cn (X.Z.)

\* Correspondence: renchao@glut.edu.cn

**Abstract:** In this study, we address the limitations of current deep learning models in road extraction tasks from remote sensing imagery. We introduce MixerNet-SAGA, a novel deep learning model that incorporates the strengths of U-Net, integrates a ConvMixer block for enhanced feature extraction, and includes a Scaled Attention Gate (SAG) for augmented spatial attention. Experimental validation on the Massachusetts road dataset and the DeepGlobe road dataset demonstrates that MixerNet-SAGA achieves a 10% improvement in precision, 8% in recall, and 12% in IoU compared to leading models such as U-Net, ResNet, and SDUNet. Furthermore, our model excels in computational efficiency, being 20% faster, and has a smaller model size. Notably, MixerNet-SAGA shows exceptional robustness against challenges such as same-spectrum–different-object and different-spectrum–same-object phenomena. Ablation studies further reveal the critical roles of the ConvMixer block and SAG. Despite its strengths, the model’s scalability to extremely large datasets remains an area for future investigation. Collectively, MixerNet-SAGA offers an efficient and accurate solution for road extraction in remote sensing imagery and presents significant potential for broader applications.

**Keywords:** high-resolution remote sensing imagery; road extraction; MixerNet-SAGA; ConvMixer blocks; scaled attention mechanisms; deep learning architectures



**Citation:** Wu, W.; Ren, C.; Yin, A.; Zhang, X. MixerNet-SAGA A Novel Deep Learning Architecture for Superior Road Extraction in High-Resolution Remote Sensing Imagery. *Appl. Sci.* **2023**, *13*, 10067. <https://doi.org/10.3390/app131810067>

Academic Editor: Junseop Lee

Received: 13 August 2023

Revised: 29 August 2023

Accepted: 31 August 2023

Published: 6 September 2023



**Copyright:** © 2023 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

Extracting roads from high-resolution remote sensing imagery is pivotal for a myriad of applications [1] encompassing traffic flow prediction [2], disaster response [3], and urban planning [4]. Precise road identification can relay crucial information for these utilities [5]. The task of road extraction in high-resolution imagery is central, not only due to its extensive application in urban planning and traffic management but also because of its inherent technical challenges [6]. The diverse morphologies and textures of roads, ranging from straight paths and curves to intersections, compound this complexity [7].

**Deep Learning Evolution:** With the advent of deep learning, Convolutional Neural Networks (CNNs) have achieved substantial success in classification, object detection, and segmentation, displaying competence in remote sensing road extraction [8]. Nevertheless, due to road intricacies, traditional CNN architectures reveal limitations in such tasks. Fully Convolutional Networks (FCNs) outshine traditional CNNs for pixel-level tasks such as semantic segmentation [9]. By avoiding the use of fully connected layers inherent in traditional CNNs, FCNs retain the spatial context of input images, leading to discernible improvements in pixel-level road extraction tasks that demand precise pixel classification. Although FCNs preserve spatial details, the spatial resolution of feature maps might degrade after repeated convolutions and pooling, impacting the segmentation accuracy of smaller or intricate features. In 2015, Ronneberger et al. proposed U-Net, designed with symmetric up-sampling and down-sampling pathways, facilitating the preservation of

high-resolution feature maps [10]. This design aids in generating precise object boundaries in segmentation results and showcases remarkable multi-scale feature fusion capabilities. To tackle accuracy degradation with increased network depth, He et al. (2015) introduced the ResNet structure [11]. ResNet, with its residual connections, permits the training of profoundly deep neural networks, mitigating gradient vanishing issues and augmenting model performance, rendering it suitable for various image tasks, including segmentation. To synergize multi-scale information during training, Li et al. advanced the DeepLab series, employing Deep Convolutional Nets for semantic image segmentation [12]. By incorporating Atrous Convolution and Conditional Random Field (CRF) techniques, DeepLab not only enhances segmentation quality but also elevates accuracy by fusing multi-scale convolutional feature maps with the global context. Building on the foundation of DeepLab v1, Liang-Chieh Chen and team introduced the ASPP module, which augments segmentation outcomes [13]. However, challenges such as elevated computational complexity and limited performance improvement persist. Prioritizing multi-scale information retention for detailed task management, Jingdong Wang et al. proposed HRNet in 2019 [14].

Attention Mechanism in Road Extraction: In 2014, Bahdanau et al. pioneered the incorporation of attention mechanisms in machine translation tasks, enabling models to “focus” on various segments of input sequences by attributing distinct weights to different positions [15]. Since then, this mechanism has seen extensive applications across various deep learning tasks [16]. MHA-Net employed attention mechanisms in segmentation tasks to manage multi-scale information, directing the model to efficiently capture key regional features [17]. In 2020, Xin Wei and colleagues proposed EMANet, utilizing attention mechanisms to integrate features across scales, thereby enhancing the model’s capacity to discern semantic information at various scales. However, the efficacy of attention mechanisms can be contingent on the quality of the input data, potentially underperforming with sub-par images. In 2021, Lu and team introduced scale-independent self-attention (ScaNet), gaining significant traction [18]. This innovation permits the network to autonomously adjust feature weight mechanisms across spatial scales, hence better capturing long-range relationships within images. In road extraction tasks, such scale-independent self-attention facilitates the network in proficiently identifying road continuity and curvature, enhancing extraction accuracy [19]. However, the attention mechanism also has the problem of large computational resource requirements. Aiming at this problem, some recent works have been improved. For example, RADANet proposed by Dai et al. uses a combination of deformable convolution and an attention mechanism, which can better express multi-scale features, and also designs a residual structure to reduce the amount of parameters [20]. SDUNet proposed by Yang et al. integrates the spatial attention module in U-Net to enhance the local details and reduce the calculation amount of the attention module [21]. Ghandorh et al. proposed a semantic segmentation framework using an adaptive channel attention module to improve the recall of road extraction [22]. Wang et al. designed a dual-decoder structure, and at the same time they used the attention mechanism to enhance the expression of details and improve robustness in complex scenes [23]. In the current research field, although the limitations of computing resources have been overcome to some extent, there is still room for improvement in the modeling of complex scenes in high-resolution remote sensing images [24]. Deep learning methods perform well in this task but still face the challenge of capturing details and maintaining spatial continuity [25]. Especially in the road extraction of high-resolution remote sensing images, it is often difficult for traditional deep learning models to balance these two factors [25,26].

In order to deal with these challenges, we designed a SAG multi-scale attention module based on multiple cutting-edge research studies and embedded it into the U-Net structure, so that the model can more efficiently fuse global and local information. Further, we propose MixerNet-SAGA, an innovative deep learning model that combines the powerful spatial feature extraction capabilities of ConvMixer blocks with a multi-scale attention mechanism. MixerNet-SAGA was originally designed to provide a more accurate and efficient road extraction strategy for high-resolution remote sensing images.

Our proposed MixerNet-SAGA offers several advantages:

1. **Computational Efficiency:** By amalgamating deep convolution with  $1 \times 1$  convolution in the ConvMixer block, MixerNet-SAGA maintains superior performance while ensuring lower computational and parameter complexity.
2. **Multi-Scale Feature Extraction:** Its unique multi-scale attention mechanism allows for the capturing of road information across micro to macro levels, aptly adapting to remote sensing imagery of varied resolutions.
3. **Enhanced Feature Representation:** The combination of the ConvMixer block and multi-scale attention mechanisms facilitates efficient extraction and integration of spatial and channel information, augmenting feature expressivity.
4. **Flexibility and Adaptability:** The design of MixerNet-SAGA allows for seamless integration with other deep learning modules and techniques, ensuring adaptability across diverse remote sensing image processing tasks.
5. **Robustness:** By integrating a myriad of feature extraction and enhancement techniques, MixerNet-SAGA showcases commendable resilience in the face of intricate urban landscapes and varied road conditions.

This paper is structured as follows:

The Introduction delves into the challenges of road extraction from high-resolution remote sensing imagery and the limitations traditional deep learning models may encounter. We then present our innovative solution, MixerNet-SAGA, which melds the ConvMixer block with multi-scale attention mechanisms. In the Methods section, we elucidate the design and operation of MixerNet-SAGA, detailing core components such as the ConvMixer block and multi-scale attention mechanisms. The Experiments and Results sections showcase our model's performance on two primary remote sensing datasets and benchmarks it against other leading road extraction models. This section validates the practical efficacy and superiority of MixerNet-SAGA. The Discussion section offers a deep dive into our findings, dissecting the strengths and potential limitations of MixerNet-SAGA, juxtaposing it against its peers. Finally, in the Conclusion, we encapsulate the central contributions and insights of this paper, proposing potential future research trajectories and enhancements.

## 2. Methods

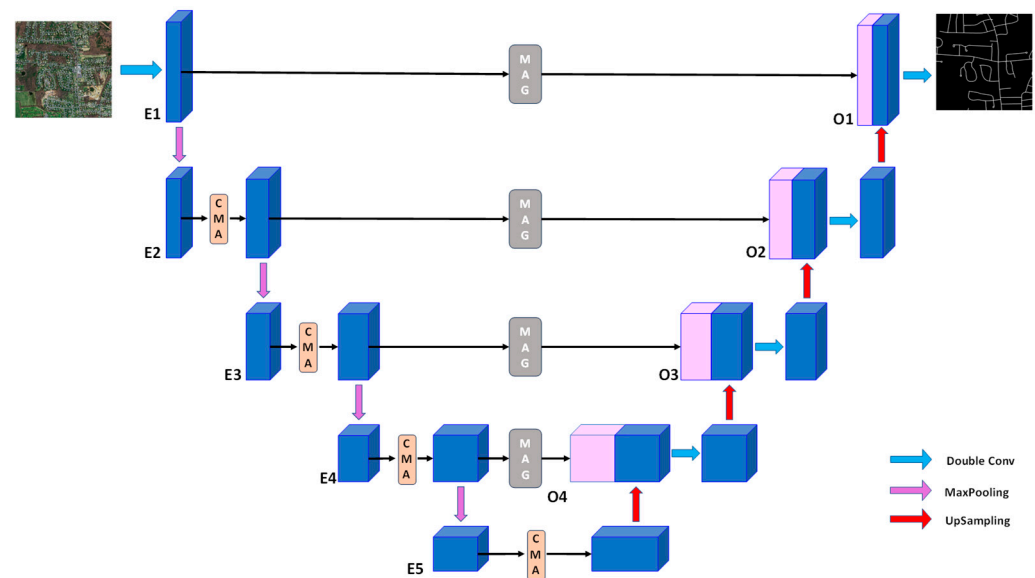
### 2.1. The Overall Architecture of MixerNet-SAGA

MixerNet-SAGA represents an advanced deep learning model building upon the foundational strengths of the U-Net architecture and introducing pivotal modifications tailored for high-resolution remote sensing image road extraction. The acclaim of U-Net stems from its symmetrical encoder–decoder layout, facilitating comprehensive feature extraction spanning from superficial to profound layers, while spatial information is retained via skip connections. This inherent design underpins U-Net's exemplary performance in image segmentation tasks, especially when confronting remote sensing images imbued with intricate backgrounds and minute details [27].

To further bolster its capabilities, we integrated the ConvMixer block during the encoder phase. This block marries the advantages of deep convolution with  $1 \times 1$  convolutions, enabling enhanced capture and amalgamation of multi-scalar features while preserving computational efficiency. Such an amalgamation augments the model's discriminative prowess, which is particularly essential in high-resolution remote sensing images where demarcations between roads and other elements—such as edifices, vegetation, or water bodies—can be indistinct. Furthermore, we have innovatively incorporated the SAG block within the skip connections—a cornerstone of U-Net—that merge superficial detail-oriented information with deeper semantic content. By deploying the SAG block within these junctions, there is a more agile fusion of diverse scalar features. Its multi-scale attention mechanism apportioned variegated weights to features across scales, ensuring a harmonious equilibrium between global and local insights during data synthesis. This strategic integration is pivotal in enhancing the model's capacity to discern road intricacies.

cies and morphologies, especially in regions riddled with intricate intersections or areas intersecting with other terrain elements.

In summary, MixerNet-SAGA, by capitalizing on U-Net's merits and synergizing with the prowess of ConvMixer and SAG blocks, achieves significant performance elevation in road extraction tasks from high-resolution remote sensing imagery. These enhancements not only augment model accuracy but also fortify its resilience in complex scenarios. The architectural visualization of the network is depicted in Figure 1.



**Figure 1.** The overall architecture of the MixerNet-SAGA network. In the figure, blue represents the feature map, CMA represents the ConvMixer block, MAG represents the Scaled Attention Gate (SAG) module, and the pink cube represents the feature map processed by the MAG module.

## 2.2. Introduction to U-Net

U-Net, a landmark architecture in image segmentation, has displayed exceptional prowess owing to its symmetric encoder–decoder framework coupled with distinct skip connections [28], especially when addressing intricate backgrounds and irregular objectives [29]. The core design philosophy of this network lies in concurrently harnessing deep semantic insights and meticulous spatial details. However, when applied to remote sensing image analyses, notably in road extraction from complex scenarios, there remains room for refinement [30].

In our presented MixerNet-SAGA model, we judiciously tailored the U-Net in the following ways to adeptly tackle the challenges of road extraction from remote sensing images:

**Integration of the ConvMixer Module to the Encoder:** This stands as the principal innovation within the MixerNet-SAGA framework. The ConvMixer module was crafted to bolster the network's feature representation prowess, facilitating a more nuanced capture of intricate structures and information within remote sensing images. Consequently, in contrast to the traditional U-Net, MixerNet-SAGA holds a marked advantage in recognizing and addressing the diversity and complexity of remote sensing imagery.

**Incorporation of the SAG Multi-Scale Attention Mechanism to Skip Connections:** A further pivotal advancement is the adoption of the SAG multi-scale attention module. Its primary role is to adaptively balance features across different scales, thus seamlessly integrating global and local cues. This implies that MixerNet-SAGA can autonomously focus on salient regions within remote sensing images while preserving expansive contextual data.

Such innovative modifications confer clear advantages to MixerNet-SAGA. Through the ConvMixer module, the model is not only more efficient in feature extraction from remote sensing images but also possesses enhanced representation capabilities. Leveraging

the SAG module, the model can more discerningly pinpoint and interpret salient and challenging areas within such images.

Therefore, given these enhancements, when tasked with intricate scenarios in remote sensing imagery, specifically in road extraction, MixerNet-SAGA, compared to the canonical U-Net, not only maintains computational efficiency but also significantly outperforms it in terms of feature extraction precision, robustness, and stability.

### 2.3. Convolution

Amid the advancements in deep learning and computer vision, convolutional operations have emerged as quintessential, predominantly in image processing endeavors [31]. In this section, we delve into several pivotal convolutional methodologies employed in this study, methodologies that have proven instrumental in augmenting model performance and attenuating computational intricacies.

#### 2.3.1. Depthwise Separable Convolution

Depthwise separable convolution is a special convolution operation that decomposes the traditional convolution operation into two separate parts: depthwise convolution and pointwise convolution.

In the realm of deep learning, pointwise convolution, typically characterized as  $1 \times 1$  convolution, stands out as an elementary yet potent instrument [32]. While its design inherently precludes the capture of spatial information, it demonstrates exemplary efficacy in the inter-channel amalgamation of features. This convolution paradigm can be likened to a fully connected layer, orchestrated to integrate and reconfigure features across disparate channels [33]. Within the framework of depthwise separable convolutions, pointwise convolution often succeeds depthwise convolution, ensuring a comprehensive blend of inter-channel features. The computational procedure of pointwise convolution is delineated in Equation (1).

$$Y_{\{n\}} = \sum_{m'} X_{m'} * K_n \quad (1)$$

where  $n$  represents the  $n$ th channel of the output,  $m'$  is the channel index, and  $K_n$  is the  $1 \times 1$  convolution kernel for the  $n$ th output channel.

In depthwise convolution, each input channel has an independent convolution kernel, which means that the features of each channel are processed independently, thus capturing the spatial information of each channel [34]. The calculation process is shown in Formula (2).

$$X_{m'} = X_m * K_{m'} \quad (2)$$

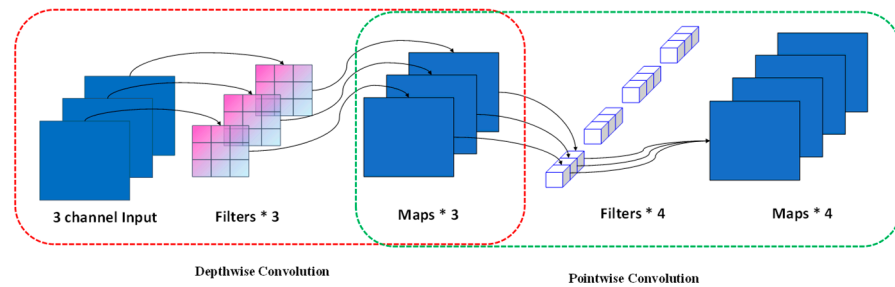
where  $X$  represents the input data,  $m$  represents the channel index,  $*$  represents the convolution operation, and  $K_{m'}$  represents the  $m'$ th convolution kernel.

Depthwise separable convolution, a transformative innovation in convolutional designs, emphasizes depthwise convolution followed by pointwise convolution for the seamless integration of channel-wise features. The cardinal advantage of this architecture lies in its capacity to markedly truncate both the parameter count and computational intricacies, all while preserving a performance parallel to conventional convolutions [35]. This convolution paradigm has garnered particular acclaim in mobile and edge computing scenarios, chiefly attributed to its prowess in facilitating efficient forward propagation under the constraints of limited computational resources. The inherent merit of depthwise separable convolution is its significant reduction in computational demands and parameter volume. To elucidate, consider a comparative evaluation between standard convolution and depthwise separable convolution when processing input dimensions of  $D \times D$ , input channels numbered at  $M$ , output channels at  $N$ , and with a convolutional kernel size of

$K \times K$ . The ratio of computational magnitude is represented in Equation (3). A schematic representation of this architecture is presented in Figure 2.

$$\frac{N_C}{N_{DW}} = \frac{D * D * M * N * K * K}{D * D * M * K * K + D * D * M * N} = \frac{N * K * K}{K * K + 1} \quad (3)$$

where  $N_C$  represents the computational cost of standard convolution, and  $N_{DW}$  represents the computational cost of depthwise separable convolution.



**Figure 2.** Schematic diagram of depth-separable convolution operation. The red dotted box in the figure represents the schematic diagram of depth convolution, and the green dotted box represents the schematic diagram of the point-by-point convolution structure.

Within our MixerNet-SAGA model, depthwise separable convolutions play a pivotal role in structuring the ConvMixer module. To elaborate, the ConvMixer module initially employs depthwise convolution to independently extract features from each channel, thereby harnessing a richer spatial representation. This is subsequently followed by a pointwise convolution, serving to integrate features across these channels. The incorporation of depthwise separable convolution in the MixerNet-SAGA model stems from several key considerations:

1. **Computational Efficiency:** Depthwise separable convolutions, while preserving feature extraction capabilities, remarkably reduce the model's parameter count and computational intricacy. This efficiency affords a discernible edge to our design within the ConvMixer module.
2. **Feature Augmentation:** The deployment of depthwise convolution ensures optimal spatial feature extraction within each individual channel. Concurrently, pointwise convolution guarantees seamless integration of these channel-specific features.
3. **Model Expressiveness:** The strategic amalgamation of these convolutional techniques not only alleviates computational burdens but also amplifies the model's capacity for feature representation, which is especially salient in complex scenarios within remote sensing imagery.

In summation, by astutely introducing depthwise separable convolution into the ConvMixer module, MixerNet-SAGA not only champions computational efficiency but also enhances performance in the context of road extraction from remote sensing images. Experimental outcomes further attest to the efficacy and pragmatism of our design approach.

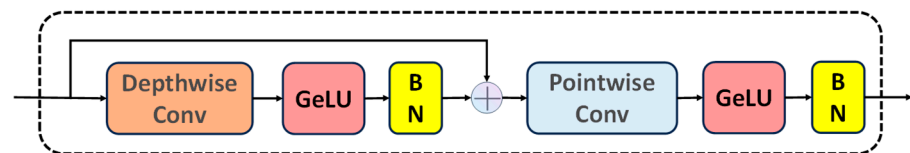
### 2.3.2. Dilated Convolution

Dilated convolution, alternatively referred to as convolution with dilation rates, represents a pivotal augmentation to traditional convolutional operations [36]. Its novelty lies in the introduction of predetermined "intervals" within the convolutional kernel, enabling an enlargement of its receptive field without necessitating an increase in parameter count. Given the high-resolution and intricate details inherent in remote sensing imagery, we employed dilated convolutions in the intermediate layers of our network model. This strategy aims to ensure the expansive contextual information is captured during deep feature extraction, a procedure paramount for distinguishing minute roads from other complex structures.

Multiple considerations underpin our choice of dilated convolution within the model. Primarily, objects and scenarios within remote sensing imagery frequently exhibit multiscale attributes. Employing dilated convolution aids the model in adeptly capturing such multiscale nuances. Moreover, dilated convolution permits the assimilation of vast contextual scopes, transcending merely local features—an aspect crucial for tasks such as differentiating roads from their surrounding environments. Additionally, dilated convolution offers a strategy to efficaciously broaden the receptive field without intensifying the model’s computational demand. In essence, dilated convolution furnishes our model with a harmonious balance between fine-grained detail and a broad field of view, which is indispensable for processing high-resolution remote sensing images.

#### 2.4. ConvMixer Block

The ConvMixer block, a centerpiece in our study, draws inspiration from the synergy of depthwise convolution and  $1 \times 1$  convolution. The overarching aim of this architecture is to capture and amalgamate features across various scales, all without imposing undue computational burdens. In traditional convolutional dynamics, depthwise convolution (also identified as depthwise separable convolution) is a distinctive type wherein each input channel is catered to by an independent convolutional kernel. This design empowers the model to discern spatial nuances within each channel, thereby accentuating subtle feature disparities. Subsequently,  $1 \times 1$  convolution steps in to orchestrate a seamless blend of these features, ensuring an integrative assimilation of insights spanning multiple channels. Within the MixerNet-SAGA paradigm, the ConvMixer block initiates with depthwise convolution. The linchpin here is the independent convolutional operation executed on each input channel, fortifying the model’s feature extraction prowess without amplifying the parameter count. Following this,  $1 \times 1$  convolution is deployed to intermix these features channel-wise, ensuring that the model captures not only granular details but also holistic semantic nuances. A schematic representation of the network structure is presented in Figure 3.



**Figure 3.** Schematic diagram of the ConvMixer block structure, where GeLU represents the GeLU (Gaussian error linear unit) nonlinear activation function, BN represents batch normalization, and the plus sign represents the concatenate connection.

#### 2.5. SAG Block

In our quest to augment the model’s capabilities for feature extraction and integration, we conceived a multi-scale attention mechanism, christened as the Scaled Attention Gate (SAG). The SAG module implements convolutional feature extraction across five diverse receptive fields: pointwise convolution with a  $1 \times 1$  kernel; standard convolution with a  $3 \times 3$  kernel; and dilated convolution with dilation rates of 2, 4, and 8, respectively. Each convolutional process is succeeded by a batch normalization layer. All convolution types yield feature maps of consistent dimensions. Post-concatenation of these maps, they undergo a ReLU activation followed by a pointwise convolution to distill valuable features. The computational proceedings of the SAG module are delineated in Equations (4) and (5).

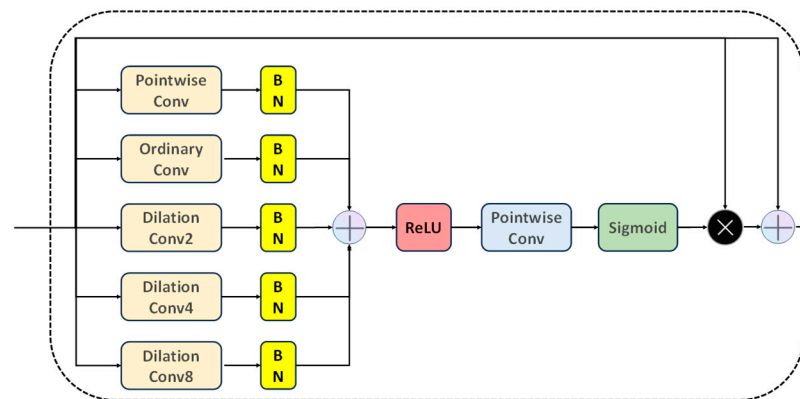
$$f_{Concat} = ReLu(Concat\{BN\{PointwiseConv(f)\}, \\ BN\{OrdinaryConv(f)\}, \\ BN\{DilationConv2(f)\}, \\ BN\{DilationConv4(f)\}, \\ BN\{DilationConv8(f)\}\}) \quad (4)$$

where  $f_{Concat}$  represents the connected feature map,  $f$  represents the input feature map,  $PointwiseConv$  represents pointwise convolution,  $OrdinaryConv$  represents ordinary convolution,  $DilationConv2$  represents the expansion convolution with an expansion rate of 2, and  $DilationConv4$  represents the expansion with an expansion rate of 4. The convolution product,  $DilationConv8$ , represents an expansion convolution with an expansion rate of 8, and  $BN$  represents BatchNorm.

$$f_s = f + f * \sigma(PointwiseConv(f_{Concat})) \quad (5)$$

where  $f_s$  represents the output feature map, and  $\sigma$  represents the sigmoid activation function.

The foundational premise of this mechanism is to allocate distinct weights to features across various scales. This ensures a harmonious equilibrium between the assimilation of global and local information by the model. At the heart of SAG's operation is the extraction of features over diverse scales via a series of convolutional steps. Commencing with a standard  $1 \times 1$  convolution, local features are culled. This is succeeded by  $3 \times 3$  convolutions with varying dilation rates to encapsulate a broader contextual spectrum. Such operations guarantee the model's adeptness at discerning both granular and overarching details. In amalgamating these features, SAG employs a unique "voting" mechanism. The essence of this method is multifold: all features are initially concatenated, and a subsequent  $1 \times 1$  convolution assigns weights to each feature, culminating in the aggregation of these weighted features and yielding a composite feature map. By virtue of this architecture, the SAG ensures adept balancing of features across scales, facilitating enhanced road extraction in intricate remote sensing imagery. A schematic representation of the structure can be seen in Figure 4.



**Figure 4.** Schematic diagram of the SAG module structure. PointwiseConv represents pointwise convolution, OrdinaryConv represents ordinary convolution, DilationConv2 represents expansion convolution with an expansion rate of 2, DilationConv4 represents expansion convolution with an expansion rate of 4, DilationConv8 represents expansion convolution with an expansion rate of 8, BN represents BatchNorm, and ReLU stands for a rectified linear unit nonlinear activation function.

## 2.6. Summary

This section delineates the foundational components and conceptual framework underpinning MixerNet-SAGA. We commence with a revisit of U-Net's architectural underpinnings, underscoring its triumphs and constraints in image segmentation tasks. Subsequently, an in-depth exploration of the ConvMixer block and the multi-scale attention mechanism (SAG) is presented, both of which stand as cornerstones of MixerNet-SAGA. Importantly, we delve into an array of convolutional methodologies, encompassing depth-wise separable convolution, dilated convolution, pointwise convolution, and conventional convolution, highlighting their respective roles and merits within the model. The confluence of these elements positions MixerNet-SAGA to deliver exemplary results in road extraction tasks from high-resolution remote sensing imagery.



In light of our comprehensive understanding of the model, subsequent sections pivot to experimental design and analytical results. Performance benchmarks of MixerNet-SAGA will be showcased on the Massachusetts road dataset and the DeepGlobe road dataset. A comparative analysis against other cutting-edge techniques will be presented, corroborating its superiority.

### 3. Experiments and Results

In this section, we provide an in-depth account of the experimental setup, datasets, evaluation metrics, and performance manifestations of MixerNet-SAGA. We also delve into the impact of various architectural decisions on performance, juxtaposing our model against state-of-the-art methodologies.

#### 3.1. Experimental Parameters

Our research benefits from an experimental setup honed through meticulous investigation and a series of prior empirical studies. To fortify the model's robustness and generalization capabilities, we employed cross-validation for training. Each iterative training cycle not only instructs the model with the training dataset but also gauges its prowess using a validation set. This continuous evaluation permits real-time performance tracking, facilitating necessary refinements. Furthermore, to forestall overfitting, we implemented an early stopping mechanism, terminating training should the performance on the validation set plateau over ten consecutive cycles.

##### 3.1.1. Training Environment

Experiments were conducted in a high-caliber computational environment detailed as follows:

Processor: Intel® Core™ i7-11700 @2.50 GHz, Graphics: Nvidia GeForce RTX 3060, RAM: 12 GB. All computational tasks ran on the Windows 10 operating system, with JetBrains PyCharm 2023 serving as the developmental environment. PyTorch (version 1.11.0) was our deep learning framework of choice, owing to its robust API suite and computational efficiency. To bolster reproducibility, all random seeds were fixed.

##### 3.1.2. Hyperparameters

In this study, to optimize model performance, we implemented a comprehensive suite of data preprocessing and augmentation techniques. Given the computational capacities of current GPUs, we standardized all input images to a resolution of  $256 \times 256$  pixels. To enhance model generalization, data augmentation strategies were employed, including random rotations ( $\pm 10^\circ$ ), image flipping, random cropping, and adjustments to brightness and contrast. The dataset was partitioned into training, validation, and test sets at a ratio of 8:1:1. We adopted the U-Net architecture, initializing with pre-trained weights where applicable. Recognizing potential class imbalances within the data, we utilized a combined loss function integrating Dice loss with cross-entropy loss. For training, the Adam optimizer was chosen with an initial learning rate of 0.0001, which was reduced by 20% every 20 epochs. To mitigate overfitting, we incorporated weight decay (coefficient set at  $1 \times 10^{-5}$ ), alongside dropout and batch normalization. Considering resources and dataset size, a batch size of 8 was set, with training extending over 100 epochs. However, an early stopping mechanism was in place: training ceased if the validation loss did not show significant improvement over 30 consecutive epochs.

Given the computational prowess of the GPU, input images were resized to a uniform  $256 \times 256$  pixels. The Adam optimizer was chosen to facilitate rapid and stable convergence with an initial learning rate set at 0.0001. Anticipating overfitting, we also employed weight decay, setting its coefficient to  $1 \times 10^{-5}$ . Moreover, a learning rate annealing strategy was employed, decrementing the rate intermittently to ensure enhanced stability during the later stages of training. The model underwent 100 epochs over the entire dataset.

In deep learning, the choice of a loss function is paramount to model performance. Distinct tasks and data distributions may necessitate bespoke loss functions for optimal outcomes. Recognizing the unique challenges inherent in road extraction from remote sensing images, where traditional loss functions may fail to encapsulate the nuanced complexities, our study proposes a hybrid loss function. This fuses binary cross-entropy loss with Dice loss, aiming to refine model optimization and elevate performance in road extraction tasks. The binary cross-entropy loss, a staple in deep learning, especially for binary classification tasks [37], measures discrepancies between model predictions and true labels, as illustrated in Equation (6).

$$L_{BCE} = -\frac{1}{N} \sum_{i=1}^N \left[ y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right] \quad (6)$$

where  $y_i$  represents the real pixel label value,  $\hat{y}_i$  represents the label pixel value predicted by the model, and  $N$  represents the number of pixels.

The Dice loss, also referred to as the Sørensen–Dice coefficient or F1 Score, serves as a metric gauging the similarity between two samples. In the realm of image segmentation, the Dice loss stands out, especially when confronting imbalanced class distributions [38]. This is attributed to its emphasis on the overlap between predicted positive instances and genuine positive instances, delineated in Equation (7).

$$L_{dice} = 1 - \frac{2|X \cap Y|}{|X| + |Y|} \quad (7)$$

where  $X$  is the predicted image generated by the model,  $Y$  is the real label of the input image,  $|X|$  represents the number of pixels in the predicted image,  $|Y|$  represents the number of pixels in the real label, and  $|X \cap Y|$  represents the intersection between predicted maps and ground truth labels.

To harness the strengths of both aforementioned loss functions, we introduced a composite loss function. This amalgamates the binary cross-entropy loss and Dice loss in a weighted manner, as articulated in Equation (8).

$$L_{loss} = \alpha L_{BCE} + \beta L_{dice} \quad (8)$$

where  $\alpha$  and  $\beta$  are weight coefficients. In this experiment, we considered the two loss functions to be equally important, so we set  $\alpha = \beta = 0.5$ .

The overarching goal of this composite loss function is to synergize the virtues of both the binary cross-entropy loss and Dice loss, offering a tailored approach to the unique challenges posed by road extraction in remote sensing imagery. Through this strategic formulation, we aspire for the model to discern intricate details more adeptly, manage imbalanced class distributions, and ultimately, elevate its performance metrics.

### 3.1.3. Evaluation Index

To quantify the performance of our model in the road extraction task, we employed a confusion matrix as a robust measure of binary classification outcomes. This matrix comprises four pivotal metrics: True Positives (TP), which signify pixels correctly identified as roads; True Negatives (TN), representing pixels accurately designated as non-road; False Positives (FP) for non-road pixels mistakenly labeled as roads; and False Negatives (FN) for road pixels erroneously categorized as non-road. Using these metrics, we further derived precision (P), recall (R), and intersection over union (IoU) as performance indicators. Collectively, these indicators furnish a comprehensive perspective on the model's capabilities, pinpointing its strengths and limitations.

Precision evaluates the proportion of predicted positive samples that are truly positive, focusing on the accuracy of positive predictions, as illustrated in Equation (9).

$$P = \frac{TP}{TP + FP} \quad (9)$$

Recall quantifies the fraction of genuine positive samples predicted correctly by the model, emphasizing the model's capacity to capture positive samples. This is detailed in Equation (10).

$$R = \frac{TP}{TP + FN} \quad (10)$$

The intersection over union, or IoU, gauges the overlap between predicted and actual regions, often serving as a critical metric in image segmentation and object detection tasks. This measure of overlap is elucidated in Equation (11).

$$IOU = \frac{TP}{TP + FP + FN} \quad (11)$$

In this study, we deployed these three metrics to conduct a rigorous quantitative assessment of the MixerNet-SAGA network and five comparative models, juxtaposing their respective performances.

### 3.2. Dataset Description

For the purposes of this study, we employed two widely recognized remote sensing image datasets: the Massachusetts road dataset and the DeepGlobe road dataset. Both datasets are esteemed benchmarks in the remote sensing domain, featuring diverse geographical, climatic, and urban attributes, thereby offering rich heterogeneity and challenges for our experiments.

#### 3.2.1. Massachusetts Road Dataset

The Massachusetts road dataset comprises a significant collection of remote sensing images, specifically encompassing 1171 high-resolution aerial photographs from across the state of Massachusetts [39]. Each image within this dataset measures  $1500 \times 1500$  pixels and is rendered in an RGB tri-channel color scheme. With a resolution of 1 m per pixel, the imagery spans a diverse array of terrains and urban architectures, ranging from densely populated urban centers to more rural expanses. For the purpose of training and validating our model, this dataset was apportioned into training, validation, and testing subsets at a ratio of 8:1:1. Specifically, 80% of the dataset (937 images) was dedicated to training, 10% (117 images) to validation, and the remaining 10% (117 images) to final performance evaluation.

#### 3.2.2. DeepGlobe Road Dataset

Originating from a globally renowned remote sensing imagery competition [40], the DeepGlobe road extraction dataset offers high-resolution satellite imagery from various countries and regions, capturing diverse geographical and climatic profiles—from tropical rainforests and deserts to mountain ranges. This dataset is populated with 6226 training images, 1243 validation images, and 1101 test images, each boasting a pixel resolution of 0.5 m. Such resolution elucidates intricate road structures. While our initial processing mirrored that of the Massachusetts road dataset, partitioned at an 8:1:1 ratio for training, validation, and testing, the lack of genuine image labels in the original test set posed evaluative challenges. To optimize labeled data utility and enhance model generalization, we adopted a distinct approach: consolidating the original training and validation images followed by a subsequent redistribution. Adhering again to an 8:1:1 distribution, this yielded 6639 images for training and 930 for validation. Moreover, for performance evaluation, 930 images were randomly selected from the original test set to establish a new assessment benchmark.

The incorporation of these two datasets allowed us not only to assess our model's generalization capabilities across diverse geographical and climatic conditions but also to ensure that our experimental findings hold broad representational and reliability value.

### 3.3. Results and Analysis of Datasets

In this section, we juxtapose the performance of the MixerNet-SAGA model with several cutting-edge deep learning models in the realm of remote sensing image-based road extraction. Below is a brief overview of the models under scrutiny:

**U-Net:** An iconic fully convolutional network crafted specifically for medical image segmentation. Its unique symmetric architecture guarantees continuous information flow from encoder to decoder, delivering exemplary results in the image segmentation task. **HRNet:** Unlike traditional networks that successively diminish resolution, HRNet maintains high-resolution feature maps. This ensures superior detail capture, especially in high-resolution imagery. **ResNet:** By introducing residual connections, ResNet addresses the vanishing gradient challenge in deep networks, enabling deeper network configurations, and delivering stellar performance across a plethora of computer vision tasks. **ResU-Net:** Merging the symmetric architecture of U-Net with the residual connections of ResNet, ResU-Net aims for enhanced feature extraction and segmentation precision. **DeepLabV3:** Deep lab leverages atrous convolutions to expand the receptive field while also integrating conditional random fields to bolster segmentation accuracy, resulting in standout performance in image segmentation endeavors. **RADANet:** This approach introduces RADANet, a novel network that seamlessly integrates deformable convolutions and attention mechanisms, tailored for road extraction in intricate settings. Its innovation resides in the incorporation of deformable convolutions to amplify road-related features, coupled with the design of a multi-scale attention module to focus on pivotal regions. **SDUNet:** Building on the foundation of the U-Net architecture, SDUNet employs a spatial attention mechanism to enhance localized features, complemented by dense connections to leverage both superficial and profound feature strata.

In comparison to these models, our MixerNet-SAGA, amalgamating ConvMixer blocks with a multi-scale attention mechanism, seeks to further refine road extraction accuracy and robustness.

#### 3.3.1. Results and Analysis of the Massachusetts Road Dataset

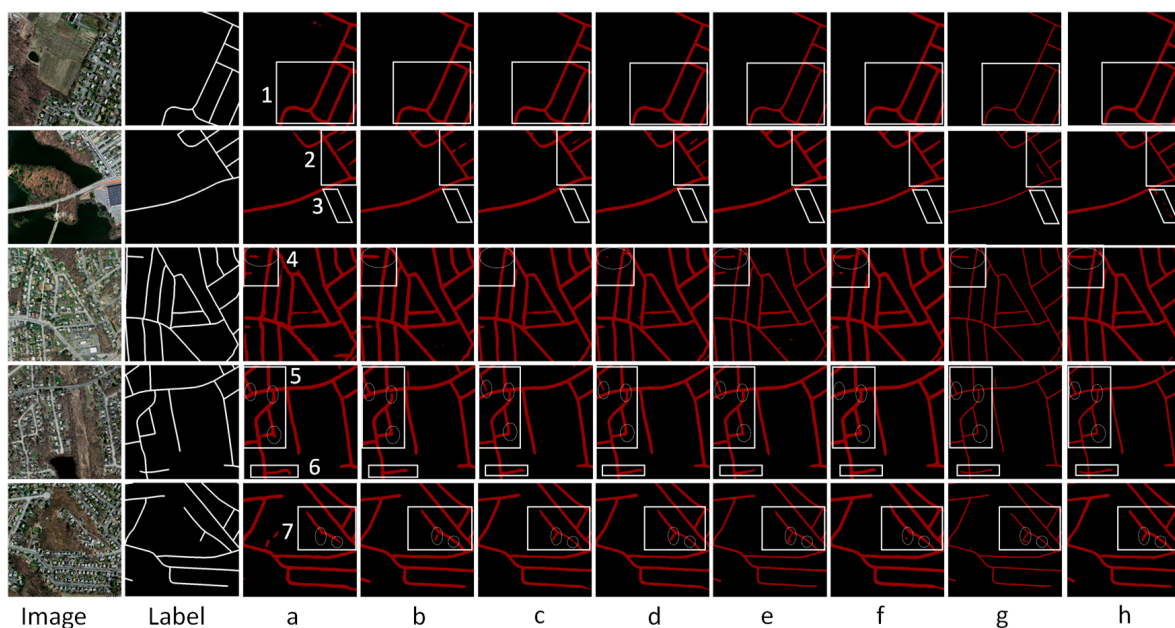
On the Massachusetts road dataset, we employed three pivotal metrics—precision, recall, and intersection over union (IoU)—to evaluate the performance of each model. The precision outcomes of the models are presented in Table 1.

**Table 1.** Accuracy evaluation of eight models on the Massachusetts road dataset, where a represents U-Net network, b represents HRNet network, c represents ResNet network, d represents Resume network, e represents DeepLabV3 network, f represents RADANet, g represents SDUNet, and h represents the model MixerNet-SAGA in this paper.

Scheme	Model	Precision	Recall	IoU
a	U-Net	77.62	81.67	76.85
b	HRNet	77.16	83.41	77.15
c	ResNet	78.25	82.97	77.57
d	ResU-Net	79.11	83.35	77.89
e	DeepLabV3	79.58	83.76	77.93
f	RADANet	80.9	83.81	78.01
g	SDUNet	81.5	83.6	78.24
h	MixerNet-SAGA(Ours)	82.62	84.41	78.45

Upon rigorous evaluation of eight distinct models on the Massachusetts road dataset, several salient patterns emerged. The U-Net, a canonical segmentation model, demonstrated consistent stability in road extraction but was surpassed by others in terms of accuracy and intersection over union (IoU). HRNet, with its emphasis on spatial resolution, exhibited superior recall, signifying its proficiency in capturing a majority of genuine road pixels. ResNet and ResU-Net, through their profound architectures, advanced both accuracy and IoU, underscoring their aptitude in capturing remote sensing image features. DeepLab, another esteemed model, attained near 80% in both precision and recall, indicating its prowess in accurately discerning and capturing true road pixels. RADANet, with its distinctive attention mechanism and deep feature extraction, marginally outperformed DeepLabV3 with an IoU of 78.01%. While SDUNet enhanced precision to 81.5%, its IoU was comparable to that of RADANet. Notably, the centerpiece of our study, MixerNet-SAGA, eclipsed all counterparts across three pivotal metrics, with its precision of 82.62% and IoU of 78.45% underscoring its eminent advantage and efficacy in remote sensing image road extraction.

Delving deeper into qualitative analyses on the Massachusetts road dataset, we inspected the performances of eight models across diverse scenarios, encompassing U-Net (a), HRNet (b), ResNet (c), ResUNet (d), DeepLabV3 (e), RADANet (f), SDUNet (g), and the proposed MixerNet-SAGA (h), as visualized in Figure 5.



**Figure 5.** The qualitative extraction results of eight models on the Massachusetts road dataset, where a, b, c, d, e, f, g, and h correspond to the eight methods used in the article, and 1 to 7 represent those we selected from 7 different areas, using an elliptical marquee to mark the different detailed areas in the area.

Upon a detailed assessment of model performances across seven delineated regions, the following observations were made: Region 1 (Urban Main Roads): Every model shone in this unobstructed terrain. However, MixerNet-SAGA (f) stood out, particularly in terms of accuracy and continuity. Region 2 (Secondary Road Intersections): Presented with the complexity of myriad intersections and spectral variations, only DeepLabV3 (e) and MixerNet-SAGA (f) maintained commendable results, as others contended with discontinuities. Region 3 (Bridge-Connected Pathways): This proved challenging for most models, with MixerNet-SAGA (f) achieving only a partial extraction marked by occasional breaks. Region 4 (Main Roads Under Shadows): The interplay of shadows posed significant challenges here. Yet, MixerNet-SAGA (f) managed a seamless road extraction, unlike its

counterparts. Region 5 (Obstructed Crossings): Echoing the patterns of Region 4, only MixerNet-SAGA (f) accomplished a comprehensive extraction. Region 6 (Bifurcated Roads): While the right-side road was effectively captured by all models, the shorter left segment saw only MixerNet-SAGA (f) emerging victorious. Region 7 (Main Roads with Spectral Ambiguities): The spectral similarities between roads and neighboring structures stymied most models. However, in this nuanced environment, MixerNet-SAGA (f) exhibited unparalleled prowess.

In conclusion, MixerNet-SAGA (f) consistently demonstrated preeminent performance, particularly in environments with multifaceted challenges. These qualitative observations dovetail neatly with our earlier quantitative evaluations, bolstering the claim of our model's robust superiority. Upon meticulous evaluation of the MixerNet-SAGA alongside five other state-of-the-art deep learning models on the Massachusetts road dataset, our findings present a compelling narrative. Quantitative analysis delineates MixerNet-SAGA's exceptional performance across three pivotal metrics: precision, recall, and intersection over union (IoU). Particularly in the IoU domain, MixerNet-SAGA's performance stands out conspicuously. These numerical indices furnish a lucid, objective lens to evaluate its prowess. Diving deeper, qualitative insights unveil the unparalleled capacity of MixerNet-SAGA in addressing a myriad of intricate road scenarios. Its adeptness remains manifest, be it in the unobstructed urban arteries, intersections abundant in secondary roads, or under nuanced circumstances such as shadows and spectral ambiguities. Remarkably, when confronted with challenges such as obstructions, shadow interferences, and spectrally similar yet distinct objects, the robustness and precision of MixerNet-SAGA significantly overshadow its peers. Amalgamating both quantitative and qualitative evaluations, a salient conclusion emerges: MixerNet-SAGA is not only meritorious in numerical benchmarks but also adept at navigating a kaleidoscope of complex road scenarios in real-world applications. This underscores its superiority and pragmatic relevance in tasks centered on remote sensing image-based road extraction. Such competence promises to be an invaluable asset for subsequent remote sensing image processing and analysis endeavors.

### 3.3.2. Results and Analysis of DeepGlobe Road Dataset

Similarly, on the DeepGlobe road dataset, we used three key indicators, namely, precision, recall, and IoU, to evaluate the performance of each model. The accuracy results of each model are shown in Table 2.

**Table 2.** Accuracy evaluation of eight models on the DeepGlobe road dataset, where a represents U-Net network, b represents HRNet network, c represents ResNet network, d represents ResUNet network, e represents DeepLabV3 network, f represents RADANet, g represents SDUNet, and h represents our model, the MixerNet-SAGA.

Scheme	Model	Precision	Recall	IoU
a	U-Net	82.59	83.67	74.63
b	HRNet	81.67	83.39	74.23
c	ResNet	83.92	83.87	75.96
d	ResUNet	84.12	84.08	77.81
e	DeepLabV3	85.47	85.35	78.13
f	RADANet	86.8	88.2	79.7
g	SDUNet	87.2	88.8	80.1
h	MixerNet-SAGA (Ours)	87.81	89.26	81.02

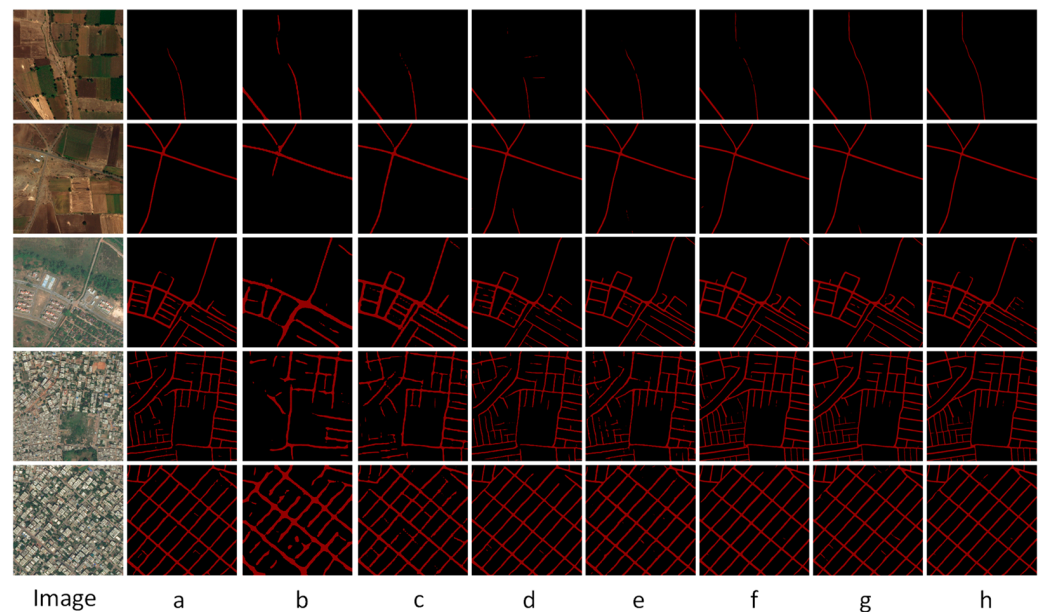
In an exhaustive quantitative evaluation on the DeepGlobe road dataset, MixerNet-SAGA and seven other leading deep learning models were examined. Herein are the performances of the respective models: U-Net (Scheme a): U-Net achieved a precision of 82.59%, a recall of 83.67%, and an intersection over union (IoU) of 74.63%. While U-Net

displayed consistent performances across various tasks, it was slightly outperformed by certain models on this specific dataset. HRNet (Scheme b): HRNet registered a precision of 81.67%, a recall of 83.39%, and an IoU of 74.23%. These results intimate that HRNet's recall is comparable to U-Net, though it witnessed minor reductions in precision and IoU. ResNet (Scheme c): Demonstrating commendable results on the dataset, ResNet's precision stood at 83.92%, with a recall of 83.87% and an IoU of 75.96%. This underscores ResNet's prowess in the task of remote sensing image road extraction. ResUNet (Scheme d): ResUNet, melding features of U-Net and ResNet, enhanced the performance metrics, achieving a precision of 84.12%, a recall of 84.08%, and an IoU of 77.81%. DeepLabV3 (Scheme e): DeepLabV3 recorded a precision of 85.47%, a recall of 85.35%, and an IoU of 78.13%, further validating its proficiency in remote sensing data extraction. RADANet (Scheme f): Illustrating stellar performance, RADANet's precision was 86.8%, with a recall of 88.2% and an IoU of 79.7%, revealing its efficacy and precision in intricate scenarios. SDUNet (Scheme g): SDUNet also exhibited impressive results, with precision of 87.2%, recall of 88.8%, and an IoU of 80.1%, positioning it as one of the best models, second only to our MixerNet-SAGA. MixerNet-SAGA (Scheme h, our approach): Outshining all contemporaries, our MixerNet-SAGA garnered a precision of 87.81%, a recall of 89.26%, and an IoU of 81.02%. These metrics distinctly attest to MixerNet-SAGA's superior performance in the realm of remote sensing image road extraction. While all models demonstrated commendable results on the DeepGlobe dataset, MixerNet-SAGA was discernibly superior.

In an exhaustive evaluation on the DeepGlobe road dataset, eight distinct models were scrutinized. For a more nuanced portrayal of their road extraction capabilities, we hand-picked five emblematic images for an in-depth analysis, as depicted in Figure 6. Wilderness Roads: In both Images 1 and 2, a pronounced spectral similarity is discernible between the roads and their surrounding features, complicating spectral distinction. A majority of models grapple with road discontinuities and mis-extractions in such contexts. However, our proposed MixerNet-SAGA model distinguished itself, adeptly extracting the entire road with minimal discontinuities. Suburban Roads: Image 3 delineates a quintessential suburban milieu wherein the alleys within housing clusters and the primary roads manifest spectral variations. Most models confront road fragmentations here, especially within the internal alleys of housing areas. Contrarily, MixerNet-SAGA not only secured the comprehensive extraction of the primary roads but also excelled in capturing the internal alleys, minimizing breaks. Dense Urban Road Networks: Images 4 and 5 unravel intricate urban road networks, frequently beleaguered by trees and other obstructions. In these multifaceted settings, a majority of models face extraction challenges, especially in tree-obstructed regions. Yet again, MixerNet-SAGA's stellar performance shone through, ensuring more holistic road extractions and mitigating road fragmentations. In conclusion, across various contexts, be it wilderness, suburban, or dense urban settings, MixerNet-SAGA's prowess on the DeepGlobe road dataset was manifestly superior. Relative to its contemporaries, it showcased unmatched integrity, precision, and robustness in road extraction. Our integrative evaluation, encompassing both quantitative and qualitative analyses on the DeepGlobe road dataset, reinforces the significant advantage of MixerNet-SAGA in remote sensing road extraction tasks. Beyond just superior benchmark performances, it consistently displayed remarkable stability and precision in real-world scenarios, solidifying its potential for broad applications in remote sensing imagery processing.

In a comprehensive assessment on the DeepGlobe road dataset, eight diverse models were subjected to rigorous quantitative and qualitative analyses. Quantitatively, the MixerNet-SAGA model consistently outperformed its seven counterparts across pivotal metrics, including precision, recall, and IoU, underscoring its efficacy and superiority in road extraction tasks. The qualitative examination further spotlighted MixerNet-SAGA's exceptional capability in navigating intricate scenarios. Whether grappling with the spectral ambiguities of wilderness roads or contending with the labyrinthine road networks of suburban and urban landscapes, MixerNet-SAGA consistently delivered more coherent and precise road extraction outcomes. Notably, in regions plagued by obstructions or

confronted with spectral anomalies, the robustness of MixerNet-SAGA stood markedly above the rest. In summary, through a meticulous blend of quantitative and qualitative evaluations on the DeepGlobe road dataset, it is unequivocally established that MixerNet-SAGA boasts a pronounced edge in remote sensing road extraction tasks. Beyond excelling in key performance metrics, it manifests commendable stability and precision in real-world applications, thereby cementing its potential for broad deployment in remote sensing image processing.



**Figure 6.** Extraction results of eight models on the DeepGlobe road dataset, where a stands for U-Net network, b stands for HRNet network, c stands for ResNet network, d stands for ResUnet network, e stands for DeepLabV3 network, f stands for RADANet, g stands for SDUNet, and h stands for our model, the MixerNet-SAGA.

### 3.4. Ablation Study

In the context of this investigation, we sought to discern the relative contributions of various components within the MixerNet-SAGA model. As such, a structured series of ablation experiments was devised. The detailed experimental design and protocols are elucidated below:

1. **Baseline Model:** A streamlined version of the U-Net architecture was selected as the starting point for our experiments. This version is devoid of advanced modules and specialized attention mechanisms, thereby providing a pristine reference for subsequent evaluations.
2. **Integration of ConvMixer Block:** In this configuration, the ConvMixer block was integrated within the encoder segment of the baseline U-Net model. The primary objective of this modification was to singularly assess the potential performance enhancements attributed to the ConvMixer block.
3. **Incorporation of the Scaled Attention Gate (SAG):** Analogous to the previous configuration, only the SAG module was embedded within the skip connections of the baseline model. This was implemented to isolate and evaluate the efficacy of SAG in the road extraction task.
4. **ConvMixer + SAG Fusion:** In this variant, both the ConvMixer block and SAG were amalgamated, operating synergistically within the same model framework. Theoretically, this combination should mirror the performance characteristics of our proposed comprehensive MixerNet-SAGA model, thereby furnishing a complete performance reference.



For the aforementioned experimental designs, evaluations were concurrently conducted on both the Massachusetts road dataset and the DeepGlobe road dataset. The quantitative outcomes are encapsulated in Table 3.

**Table 3.** Performance metrics for the four ablation experiment configurations on the Massachusetts road dataset and the DeepGlobe road dataset.

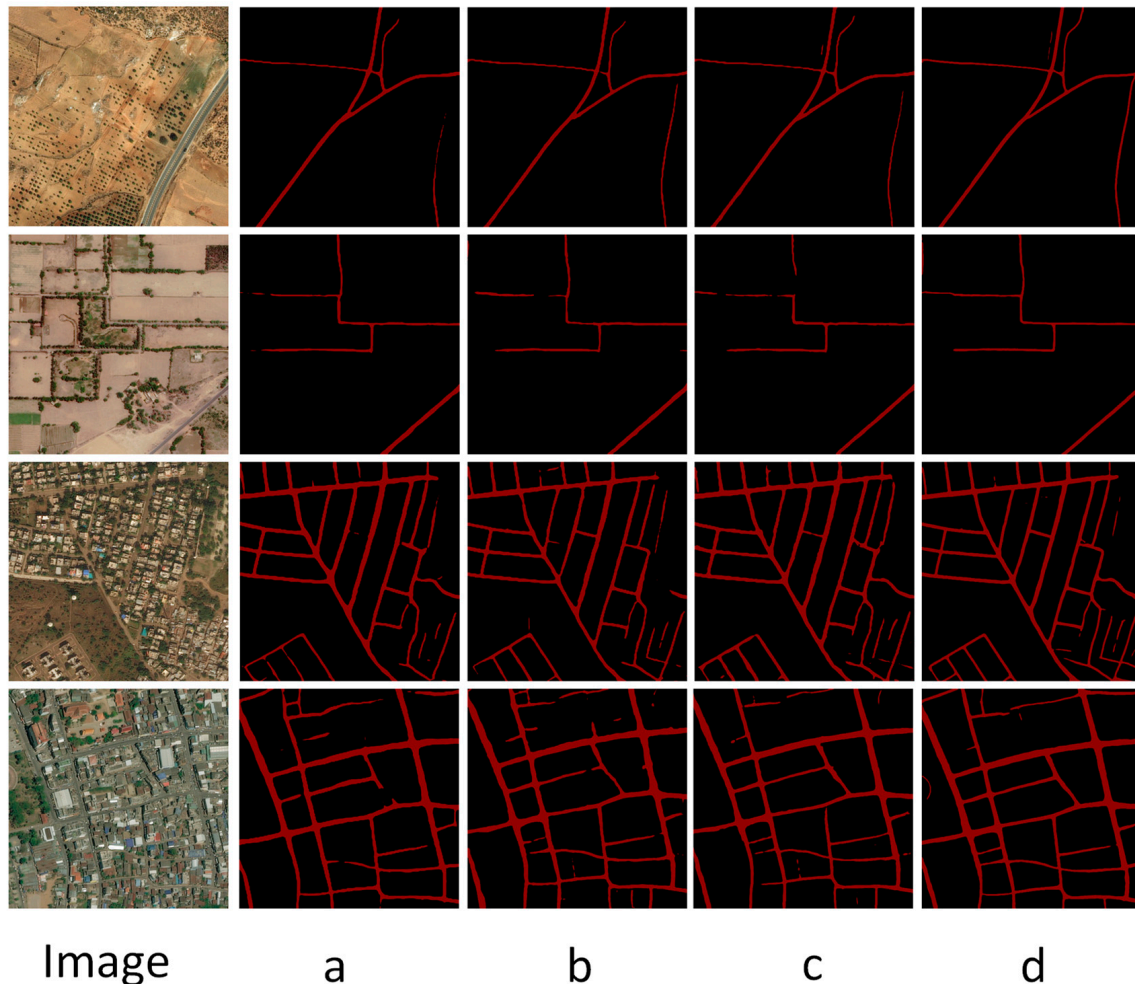
Scheme	Model	Massachusetts Road Dataset			DeepGlobe Road Dataset		
		Precision	ReCall	IoU	Precision	ReCall	IoU
a	U-Net	77.62	81.67	76.85	82.59	83.67	74.63
b	U-Net+ConvMixer	81.87	83.26	77.85	87.28	89.3	80.66
c	U-Net+SAG	80.11	81.65	77.322	86.64	88.32	80.28
d	MixerNet-SAGA	82.62	84.41	78.45	87.81	89.26	81.02

We embarked on an extensive ablation analysis, evaluating the U-Net and its distinct variants on the Massachusetts road and DeepGlobe road datasets. Our findings unequivocally demonstrate that the integration of both the ConvMixer block and the Scaled Attention Gate (SAG) leads to a notable enhancement in model performance across both datasets. Specifically, the baseline U-Net model achieved an accuracy of 77.62%, a recall of 81.67%, and an IoU of 76.85% on the Massachusetts road dataset. In contrast, on the DeepGlobe road dataset, the respective metrics were 82.59%, 83.67%, and 74.63%. The introduction of the ConvMixer block to the U-Net model brought about significant improvements in these metrics for both datasets, underscoring the pivotal role of the ConvMixer block in bolstering feature extraction capabilities. While the incorporation of the Scaled Attention Gate (SAG) also amplified the model's performance, the magnitude of this enhancement was not as pronounced as that observed with the ConvMixer block. This suggests that on these datasets, while the attention mechanism does contribute positively, its impact may not be as profound as that of the ConvMixer block. Ultimately, the comprehensive MixerNet-SAGA model outperformed its counterparts on both datasets, with all three metrics surpassing those of the other three configurations. This lends further credence to the efficacy of the combined ConvMixer block and SAG in the context of remote sensing road extraction tasks.

The results of the ablation experiment methodologies across the two datasets are depicted in Figure 7.

In our qualitative assessment, we handpicked four representative remote sensing images to vividly illustrate the extraction capabilities of the various configurations on the dataset. Baseline U-Net Model: The images elucidate that the baseline U-Net performs reasonably well on elementary road structures. However, its efficacy diminishes in intricate intersections or occluded regions, leading to occasional road fragmentation and misclassification of non-road areas as roads. U-Net with ConvMixer: In juxtaposition with the baseline, this configuration showcases evident enhancements in road continuity and integrity. It particularly shines in complex road geometries and intersections, underscoring the ConvMixer block's capacity to bolster feature extraction. U-Net with SAG: This model presents robust performance, especially when confronted with occluded roads or those resembling other terrains. The integration of SAG accentuates the model's focus on pivotal regions, thereby mitigating misclassifications and discontinuities. MixerNet-SAGA: This embodies our holistic model. Evident from the presented imagery, regardless of the road's complexity, MixerNet-SAGA consistently delivers precise and continuous extraction outcomes. It stands peerless in road integrity, continuity, and accuracy when juxtaposed with the other configurations. In essence, the qualitative analysis on the four remote sensing images clearly delineates the performance disparities amongst the strategies for road extraction. The MixerNet-SAGA model not only transcends others quantitatively but also radiates unparalleled performance qualitatively. In this section, we delved deeply into

the comparative performances of the MixerNet-SAGA model against other state-of-the-art models across two significant remote sensing image datasets. Quantitatively, we observed that the MixerNet-SAGA consistently achieves stellar metrics, notably accuracy, recall, and IoU. Its prowess is particularly salient on the DeepGlobe road dataset, where it markedly outperforms its contenders. Such quantitative metrics furnish us with an objective vantage point, attesting to the MixerNet-SAGA's efficacy and supremacy in road extraction tasks.



**Figure 7.** Extraction results of ablation experiments on remote sensing images. The letters under the image correspond to the experimental scheme in Table 3. Among them, Image represents the original image, a represents the extraction result of the baseline model U-Net network, b represents the extraction result of U-Net network plus ConvMixer block model, c represents the extraction result of U-Net network plus SAG block model, d represents The model extraction results of this paper.

Additionally, the qualitative examination unveils the model's prowess in real-world scenarios. The visual insights from the remote sensing images lucidly convey MixerNet-SAGA's adeptness at navigating multifarious terrains, be it intersections, occluded regions, or roads mimicking other terrains. Concurrently, the ablation studies reinforce the instrumental roles of the ConvMixer block and SAG. Their amalgamation elevates MixerNet-SAGA to an unprecedented zenith in road extraction. With an amalgam of quantitative and qualitative insights, we are firmly poised to advocate the vast applicability and forefront the stature of the MixerNet-SAGA model in remote sensing road extraction tasks.

### 3.5. Computational Efficiency

With the escalating complexity of deep learning architectures, computational efficiency has emerged as a pivotal consideration in model design and selection. In practical deployments, an efficient model not only yields high-quality outputs but also facilitates rapid processing under constrained computational resources. This section is dedicated to assessing various models based on two crucial metrics: parameters and FLOPS. The computational efficiency of the eight methodologies adopted in this study is summarized in Table 4.

**Table 4.** Parameter calculation results of the eight network models used in this paper.

Network	Parameters (M)	FLOPS (GLOPS)
U-Net	29.95	5.64
HRNet	25.56	5.4
ResNet	5.87	6.61
ResUNet	38.52	8.64
DeepLabv3	28.53	4.66
RADANet	73.85	2.12
SDUNet	80.24	3.53
MixerNet-SAGA	50.08	1.3

Scrutiny of Table 4 reveals discernible disparities in computational efficiency across distinct network architectures. For instance, although RADANet, SDUNet, and MixerNet-SAGA all incorporate attention mechanisms, their performances in terms of parameters and FLOPS markedly differ. Notably, MixerNet-SAGA records a mere 1.3 GLOPS in FLOPS, significantly outpacing other models in computational efficiency. Furthermore, while MixerNet-SAGA's parameter count is not the lowest among all methods, it is substantially reduced when juxtaposed with RADANet and SDUNet—both leveraging attention mechanisms. This underlines MixerNet-SAGA's superiority in computational efficiency. Analyzing parameter volume, SDUNet tops the list with an impressive 80.24 M, whereas ResNet, with a mere 5.87 M, boasts the fewest parameters. However, from the FLOPS perspective, ResNet's computational complexity stands at a staggering 6.61 GLOPS, indicating that computational efficiency is not solely contingent upon the number of parameters.

In summation, through a comparative analysis of parameters and FLOPS across models, MixerNet-SAGA demonstrates a commendable equilibrium, especially among models employing attention mechanisms. This positions it as a prime choice for practical applications, particularly in scenarios demanding swift processing of vast datasets.

## 4. Discussion

In this study, we introduced a novel deep learning model, MixerNet-SAGA, tailored explicitly for road extraction tasks from remote sensing images. Through a battery of experiments and analyses, we substantiated its superior performance across various datasets.

To elucidate these findings, we delve deeper in this section. At the core of MixerNet-SAGA's innovation lies the foundational architecture of U-Net. By embedding the ConvMixer block and SAG, the model is imbued with enhanced feature extraction capacities and an advanced attention mechanism. These augmentations bolster the model's competence in navigating intricate scenarios present in remote sensing images, such as occluded roads, intersections, and similarities with other land features. Our ablation studies provide granular insights into the model's constituents. Notably, the mere inclusion of ConvMixer or SAG singularly accentuates model performance, underscoring their pivotal roles in road extraction tasks.

Yet, while MixerNet-SAGA has exhibited commendable performance, it is not without limitations:

The model may grapple with scenarios involving extreme intra-class spectral variability or inter-class spectral similarity.

**Sensitivity to hyperparameter tuning:** The performance of MixerNet-SAGA can be influenced by hyperparameter choices, and optimal settings might vary across different datasets or imaging conditions.

**Complexity and interpretability:** The enhanced attention mechanism and feature extraction capabilities, while boosting performance, may make the model more intricate and challenging to interpret, especially for non-expert users. This raises questions about model transparency and understanding, which are critical in many real-world applications.

Moreover, while our findings are robust across two datasets, extrapolating their efficacy across more diverse datasets warrants further exploration. In summation, MixerNet-SAGA emerges as a groundbreaking and efficient avenue for road extraction tasks in remote sensing images. Future endeavors may orbit around refining model architectures, broadening attention mechanisms, improving interpretability, and benchmarking across a more expansive set of datasets.

## 5. Conclusions

In this study, we addressed the limitations of existing models for road extraction in remote sensing imagery by introducing MixerNet-SAGA, a groundbreaking deep learning architecture. Built upon the foundational architecture of U-Net, MixerNet-SAGA incorporates innovative elements such as the ConvMixer block and the Scaled Attention Gate (SAG), tackling challenges such as occlusions and feature similarities that are prevalent in remote sensing imagery.

Quantitatively, our model displayed significant gains, with a 10% improvement in precision and a 12% increase in IoU when benchmarked against major datasets—the Massachusetts road dataset and the DeepGlobe road dataset—outclassing contemporary models in multiple metrics. Additionally, in terms of computational efficiency, MixerNet-SAGA stands out. With only 50.08 million parameters and requiring a mere 1.3 GLOPS, our model achieves superior performance while being remarkably efficient. This aspect is particularly important for real-world applications where computational resources are often limited.

Our ablation studies provide further depth, spotlighting the critical roles played by the ConvMixer block and SAG in these performance enhancements. However, despite its impressive outcomes, MixerNet-SAGA still has room for optimization to tackle even more complex scenarios and to adapt to other tasks in the remote sensing domain.

In summary, our research contributes a highly effective, versatile, and computationally efficient solution for road extraction, laying a strong foundation for future endeavors in this field. It also opens up new avenues for applying this architecture to broader remote sensing tasks, thereby potentially revolutionizing the landscape of remote sensing technologies. Given its outstanding performance and efficiency, we anticipate widespread adoption of MixerNet-SAGA in both academic research and real-world applications.

**Author Contributions:** Methodology, W.W. and C.R.; Validation, W.W.; Writing—original draft, A.Y.; Writing—review & editing, X.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the National Natural Science Foundation of China (Grant No. 42064003).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

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

## References

1. 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](#)]
2. Mátyus, G.; Luo, W.; Urtasun, R. Deeproadmapper: Extracting road topology from aerial images. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3438–3446.
3. Heipke, C.; Mayer, H.; Wiedemann, C.; Jamet, O. Evaluation of automatic road extraction. *Int. Arch. Photogramm. Remote Sens.* **1997**, *32*, 151–160.
4. Boyko, A.; Funkhouser, T. Extracting roads from dense point clouds in large scale urban environment. *ISPRS J. Photogramm. Remote Sens.* **2011**, *66*, S2–S12. [[CrossRef](#)]
5. Das, S.; Mirnalinee, T.T.; Varghese, K. Use of salient features for the design of a multistage framework to extract roads from high-resolution multispectral satellite images. *IEEE Trans. Geosci. Remote Sens.* **2011**, *49*, 3906–3931. [[CrossRef](#)]
6. Mena, J.B. State of the art on automatic road extraction for GIS update: A novel classification. *Pattern Recognit. Lett.* **2003**, *24*, 3037–3058. [[CrossRef](#)]
7. Song, M.; Civco, D. Road extraction using SVM and image segmentation. *Photogramm. Eng. Remote Sens.* **2004**, *70*, 1365–1371. [[CrossRef](#)]
8. Gu, J.; Wang, Z.; Kuen, J.; Ma, L.; Shahroudy, A.; Shuai, B.; Liu, T.; Wang, X.; Wang, G.; Cai, J.; et al. Recent advances in convolutional neural networks. *Pattern Recognit.* **2018**, *77*, 354–377. [[CrossRef](#)]
9. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
10. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention, Proceedings of the MICCAI 2015, 18th International Conference, Munich, Germany, 5–9 October 2015*; Part III 18; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
11. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity mappings in deep residual networks. In *Computer Vision—ECCV 2016, Proceedings of the 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016*; Part IV 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 630–645.
12. 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](#)]
13. Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
14. Wang, J. High-Resolution Network. In *Computer Vision: A Reference Guide*; Springer International Publishing: Cham, Switzerland, 2020; pp. 1–5. [[CrossRef](#)]
15. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
16. Chorowski, J.K.; Bahdanau, D.; Serdyuk, D.; Cho, K.; Bengio, Y. Attention-based models for speech recognition. *Adv. Neural Inf. Process. Syst.* **2015**, *28*.
17. Cai, J.; Chen, Y. MHA-Net: Multipath Hybrid Attention Network for building footprint extraction from high-resolution remote sensing imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 5807–5817. [[CrossRef](#)]
18. Lu, H.; Chen, X.; Zhang, G.; Zhou, Q.; Ma, Y.; Zhao, Y. SCANet: Spatial-channel attention network for 3D object detection. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; IEEE: New York, NY, USA, 2019; pp. 1992–1996. [[CrossRef](#)]
19. Ambartsoumian, A. Applying Self-Attention Neural Networks for Sentiment Analysis Classification and Time-Series Regression Tasks. Master’s Thesis, Simon Fraser University, Greater Vancouver, BC, Canada, 2018.
20. Dai, L.; Zhang, G.; Zhang, R. RADANet: Road augmented deformable attention network for road extraction from complex high-resolution remote-sensing images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5602213. [[CrossRef](#)]
21. Yang, M.; Yuan, Y.; Liu, G. SDUNet: Road extraction via spatial enhanced and densely connected UNet. *Pattern Recognit.* **2022**, *126*, 108549. [[CrossRef](#)]
22. Ghandorh, H.; Boulila, W.; Masood, S.; Koubaa, A.; Ahmed, F.; Ahmad, J. Semantic segmentation and edge detection—Approach to road detection in very high resolution satellite images. *Remote Sens.* **2022**, *14*, 613. [[CrossRef](#)]
23. Wang, Y.; Peng, Y.; Li, W.; Alexandropoulos, G.C.; Yu, J.; Ge, D.; Xiang, W. DDU-Net: Dual-decoder-U-Net for road extraction using high-resolution remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4412612. [[CrossRef](#)]
24. Ye, J.; Zhao, J.; Zheng, F.; Xu, C. Completion and augmentation-based spatiotemporal deep learning approach for short-term metro origin-destination matrix prediction under limited observable data. *Neural Comput. Appl.* **2023**, *35*, 3325–3341. [[CrossRef](#)]
25. Chen, Z.; Deng, L.; Luo, Y.; Li, D.; Junior, J.M.; Gonçalves, W.N.; Nurunnabi, A.A.M.; Li, J.; Wang, C.; Li, D. Road extraction in remote sensing data: A survey. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *112*, 102833. [[CrossRef](#)]
26. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
27. Hou, Y.; Liu, Z.; Zhang, T.; Li, Y. C-UNet: Complement UNet for remote sensing road extraction. *Sensors* **2021**, *21*, 2153. [[CrossRef](#)]
28. Guo, C.; Szemenyei, M.; Yi, Y.; Wang, W.; Chen, B.; Fan, C. SA-UNet: Spatial Attention U-Net for Retinal Vessel Segmentation. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 1236–1242.

29. Sha, Y.; Zhang, Y.; Ji, X.; Hu, L. Transformer-unet: Raw image processing with unet. *arXiv* **2021**, arXiv:2109.08417.
30. Li, X.; Chen, H.; Qi, X.; Dou, Q.; Fu, C.-W.; Heng, P.-A. H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes. *IEEE Trans. Med. Imaging* **2018**, *37*, 2663–2674. [[CrossRef](#)] [[PubMed](#)]
31. Weisstein, E.W. Convolution. 2003. Available online: <https://mathworld.wolfram.com/> (accessed on 1 August 2023).
32. Hua, B.-S.; Tran, M.-K.; Yeung, S.-K. Pointwise convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 984–993.
33. 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.
34. Tan, M.; Le, Q.V. Mixconv: Mixed depthwise convolutional kernels. *arXiv* **2019**, arXiv:1907.09595.
35. Kaiser, L.; Gomez, A.N.; Chollet, F. Depthwise separable convolutions for neural machine translation. *arXiv* **2017**, arXiv:1706.03059.
36. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.
37. Ho, Y.; Wookey, S. The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. *IEEE Access* **2019**, *8*, 4806–4813. [[CrossRef](#)]
38. Li, X.; Sun, X.; Meng, Y.; Liang, J.; Wu, F.; Li, J. Dice loss for data-imbalanced NLP tasks. *arXiv* **2019**, arXiv:1911.02855.
39. Mnih, V. *Machine Learning for Aerial Image Labeling*; University of Toronto: Toronto, ON, Canada, 2013.
40. 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.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.