



Article

CE-RoadNet: A Cascaded Efficient Road Network for Road Extraction from High-Resolution Satellite Images

Ke-Nan Cheng ¹, Weiping Ni ^{1,*}, Han Zhang ¹, Junzheng Wu ¹, Xiao Xiao ¹ and Zhigang Yang ²

¹ Northwest Institute of Nuclear Technology, Xi'an 710024, China; chengkenan@nint.ac.cn (K.-N.C.); zhanghan@nint.ac.cn (H.Z.); wujunzheng@nint.ac.cn (J.W.); xiaoxiao@nint.ac.cn (X.X.)

² School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China; zgyang@mail.nwpu.edu.cn

* Correspondence: niweiping@nint.ac.cn

Abstract: The reconstruction of road networks from high-resolution satellite images is of significant importance across a range of disciplines, including traffic management, vehicle navigation and urban planning. However, existing models are computationally demanding and memory-intensive due to their high model complexity, rendering them impractical in many real-world applications. In this work, we present Cascaded Efficient Road Network (CE-RoadNet), a novel neural network architecture which emphasizes the elegance and simplicity of its design, while also retaining a noteworthy level of performance in road extraction tasks. First, a simple encoder–decoder architecture (Effi-RoadNet) is proposed, which leverages smoothed dilated convolutions combined with an attention-guided feature fusion module to aggregate features from multiple levels. Subsequently, an extended variant termed CE-RoadNet is designed in a cascaded architecture to enhance the feature representation ability of the model. Benefiting from the concise network design and the prominent representational ability of the stacking mechanism, our network can accomplish better trade-offs between accuracy and efficiency. Extensive experiments on public road datasets demonstrate that our approach achieves state-of-the-art results with lower complexity. All codes and models will be released soon to facilitate reproduction of our results.



Academic Editors: Liangxiu Han, Stefano Pignatti, Jiali Shang, Wenjiang Huang and Yanbo Huang

Received: 8 January 2025

Revised: 22 February 2025

Accepted: 24 February 2025

Published: 27 February 2025

Citation: Cheng, K.-N.; Ni, W.; Zhang, H.; Wu, J.; Xiao, X.; Yang, Z. CE-RoadNet: A Cascaded Efficient Road Network for Road Extraction from High-Resolution Satellite Images. *Remote Sens.* **2025**, *17*, 831. <https://doi.org/10.3390/rs17050831>

Copyright: © 2025 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/>).

Keywords: road extraction; remote sensing image; encoder–decoder

1. Introduction

Automatic road extraction from remote sensing images plays an important role in road network planning, traffic control and management, etc. Compared to the traditional algorithms that focus on hand-crafted features to extract roads [1–3], deep-learning-based models [4–6] have achieved a satisfactory accuracy in the task of automatic road extraction from remote sensing imagery [7,8] in an economic and effective way. There are two main deep learning strategies for this task: (1) graph-based approaches [9–17], and (2) segmentation-based approaches [18–22].

Graph-based methods have gained popularity recently, which could enhance the topological correctness of roads and consequently improve the completeness of predicted road results. In comparison to methods that use intermediate representations like mask images, they directly predict graph nodes and edges in a vectorized form. An autoregressive approach is utilized in graph-based methods to construct road graphs incrementally, which, however, makes it difficult to parallelize as this approach relies on the outcomes of previous steps. Although keypoint-based graph representations have exhibited an outstanding

performance, many methods of this type face a considerable drawback in terms of the substantial training cost. This high training cost can be traced back to the oversampling strategy which is utilized in defining the so-called “key points” in the graph annotations.

The segmentation-based methods are characterized by the classification of each pixel in the images by drawing different colors to different categories. These methods can be categorized into two primary types: those based on Transformers and those based on convolutional neural networks (CNNs).

Since the emergence of Vision Transformers (ViTs) [23,24], they have demonstrated remarkable proficiency in a diverse array of vision tasks and have contributed pioneering concepts for the design of vision networks, so the application of ViTs in road extraction tasks has become pervasive. Due to the self-attention mechanism, ViTs differ from traditional convolutional networks in their capacity for long-range context modeling. However, in contrast to natural image segmentation tasks, road extraction tasks require the model to pay more attention to local or relatively large receptive fields rather than the global context information. As shown in Figure 1, in the context of natural images, a large receptive field is a requisite for recognizing the horse and the girl. Conversely, in the domain of remote sensing images, road detection merely necessitates the utilization of a local window. Some works combine CNNs and Transformers to simultaneously capture local and global interactions. C²net [25] adopted the pre-trained ResNet34 as the encoder, and Swin-Transformer blocks were used in the sub-module in order to merge the convolution and the Transformer network to utilize the advantages of both networks. However, it is essential to acknowledge that the implementation of shifted window self-attention in Swin-Transformer demands sophisticated engineering, which includes the rewriting of CUDA codes.



Figure 1. Natural image and remote sensing image segmentation task.

Most recently, it has been demonstrated in several studies [26–28] that a pure CNN can attain satisfactory results as well as ViTs and graph-based methods [25]. Firstly, CNNs inherently excel at capturing local spatial patterns and hierarchical features through the convolutional layers, which is crucial for delineating intricate road networks from high-resolution imagery. This localized feature extraction capability allows CNNs to effectively discern fine-grained details such as road edges, textures and connectivity, which are often challenging for graph-based methods that rely on the predefined graph structures and may struggle with capturing subtle spatial variations. Secondly, CNNs are more computationally

efficient during inference compared to ViTs, since ViTs require substantial computational resources due to their self-attention mechanisms, making them less practical for large-scale, high-resolution datasets. Additionally, CNNs benefit from extensive architectural optimizations, enabling them to achieve robust performance with relatively fewer data requirements compared to ViTs and graph-based approaches. Overall, CNNs offer a balanced combination of precision, efficiency and scalability, making them particularly well suited for high-resolution road extraction tasks.

Drawing inspiration from the SOTA CNN models [25,26,28], a range of progressively intricate approaches based on increasingly sophisticated CNN architectures have been steadily enhancing their performance on well-established benchmark road extraction datasets. However, these methods that rely on dense connectivity cannot realize the effective interaction of full-scale features. In addition, these models require high-performance devices for training, which limit their practical applications.

In this paper, we adopt a reflective stance and delve into the actual requirements underlying the complexity of current models. To the best of our knowledge, current road detection datasets [29,30] contain thousands of image pairs, which are far from sufficient compared to natural image datasets [31,32]. For datasets like ImageNet [32], the scale of the images is often limited to an individual or a small number of objects, where global context is more important for feature representation. In contrast, in remote sensing imagery, the scale of each object is generally smaller, making the acquisition of global features less meaningful. Instead, local features become more critical, which is why CNN models offer a significant advantage over Transformer-based models on remote sensing imagery processing tasks. Due to the translation equivariance and locality characteristics, CNNs can generalize better when trained on insufficient amounts of data like that in road detection datasets.

In light of the fact that current models are confronted with challenges pertaining to hefty computational demands, it is imperative to design lightweight networks without compromising high accuracy. Considering the special features of the task, the model should have a relatively large receptive field while preserving the detailed spatial information. To this end, we propose a simplified approach that employs a single convolutional layer in lieu of the traditional convolutional blocks in classic CNN architectures. To ensure the preservation of fine-grained road details and mitigate information loss during downsampling operations, we only perform downsampling twice, while four downsampling layers are adopted in conventional models. However, a pivotal challenge arises: the reduced downsampling operation results in a smaller receptive field and a higher number of FLOPs of the model, making it difficult to capture the elongated distribution characteristics of roads. To address this limitation, we incorporate dilated convolution that can enlarge the receptive field without sacrificing the spatial resolution. Despite its advantages, it could introduce gridding artifacts, leading to a detrimental effect on the performance of CNNs. To address these issues, we draw inspiration from the concept of smoothed dilated convolution [33], stacking this block six times and effectively resolving the aforementioned problems. In addition, to fuse the different levels of features, we further propose an attention-guided feature fusion module to determine the importance of features of different levels and aggregate them based on their corresponding importance maps. Only a single convolutional layer is used to learn the attention maps of different levels and feature fusion is processed via element-wise multiplication. In this way, a simple encoder–decoder architecture is constructed, dubbed as Effi-RoadNet. Since the encoder of the network contains only 17 convolutional layers, which is lightweight but might suggest a limited representational ability of the model, we cascade this encoder–decoder architecture twice to enhance its feature representation ability with an acceptable increase in complexity. The encouraging results obtained by the model on two benchmark datasets demonstrate its efficacy, concur-

rently ensuring that the complexity remains at a minimal level. The contributions can be summarized as follows:

- (1) We introduce a simple yet effective Effi-RoadNet, which utilizes smoothed dilated convolution to mitigate gridding artifacts and incorporates an attention-guided feature fusion subnetwork to aggregate features from different levels.
- (2) We propose a straightforward extension, dubbed as CE-RoadNet, which further improves the feature representation ability for effective road network detection, and achieves superior performance on established benchmark datasets.
- (3) We evaluate our CE-RoadNet against baseline methods on two large-scale public datasets, demonstrating its superior performance.

The rest of the contents of this article are structured as follows: Section 2 introduces the related works about road extraction tasks. Section 3 provides a comprehensive description of CE-RoadNet, including the design of the attention-guided feature fusion module and Effi-RoadNet. Section 4 showcases the experiments validating the effectiveness and generalization capability of the proposed approach across various datasets. Section 5 presents our conclusions.

2. Related Works

2.1. Road Extraction Using ViTs

Roads are widely distributed and slender in remote sensing images. This characteristic often renders traditional methods deficient in terms of global context perception and localization accuracy. The Transformer architecture has demonstrated remarkable proficiency in capturing global information and making use of contextual cues from the input imagery. In recent years, a plethora of methods based on Vision Transformers have been put forward for road extraction tasks. Nevertheless, only a handful of them rely solely on ViTs, since they lack the spatial inductive bias and demand excessive data and computational resources.

By integrating ViTs into a CNN architecture, most methods enable the Transformer to interact with other components, thereby collaboratively accomplishing the task of road extraction [13,34–37]. Ge et al. [34] integrated the Swin-Transformer, which is known for its efficient multi-head and shifted window self-attention, into a U-shaped architecture to boost global learning. The Swin-Transformer was utilized in TransRoadNet [35] to downsample the feature maps. A Transformer-based approach, presented by Zhang et al. [36], comes with modules dedicated to the extraction of detailed road features as well as the fusion of global/local contexts. Yang et al. [37] proposed SSEANet, in which the CNN and Swin-Transformer are jointly trained with consistency loss to enhance their cross-supervised capabilities. ResUNetFormer [38] proposed a deep semantic segmentation neural network that utilizes the abilities of residual learning, HetConvs, UNet and vision transformers. Hetang et al. [9] made use of the intrinsic strengths of the Segment Anything Model, which adopts Transformers as a special form of graph convolutional network. A ViT-B SAM model is used as the backbone, and two decoders are adopted to predict the road map, vertex and road topology. RNGDet++ [13] employed a DETR-like Transformer network to fully utilize its multi-scale features. Training this model on a moderately sized dataset consisting of thousands of images demands approximately 192 GPU hours. In general, the combination of self-attention and Conv-blocks can strike a balance between efficiency and effectiveness, but these approaches still need a substantial training cost.

Transformer-based models exhibit several disadvantages in road extraction. Firstly, they possess high computational complexity, entailing long training/inference times and stringent hardware requirements. Secondly, they rely heavily on large-scale, high-quality datasets, since insufficient or poor data can trigger overfitting and limit generalization. Thirdly, while being good at capturing long-distance dependencies, these models are weak

in delineating local road features precisely. Finally, their intricate structures result in poor interpretability, complicating debugging and result evaluation.

2.2. Graph-Based Neural Architectures

To better preserve the connectivity of roads, some methods perform preprocessing on ground truth road maps. They convert these images into graph-structured data consisting of nodes and edges, which can better showcase the linear connectivity characteristics of roads [10–17]. Batra et al. [14] propose a multi-branch CNN model to jointly learn orientation and per-pixel segmentation, thereby improving road connectivity. They also design a connectivity refinement mechanism that closes the gaps in the prediction. Vecrood [16] employ a point-based iterative graph exploration scheme. The core of this method lies in maintaining the precise alignment between the graph and the actual roads through segmentation-cue guidance and flexible step sizes. Zao et al. [17] predict vectorized representations by employing decoupled lightweight network branches. This method circumvents the need for a computationally intensive tracking framework or the use of heavy graph tensors, thereby enhancing efficiency and performance. Xu et al. [39] present a novel approach to computing vector road maps from satellite remotely sensed images, building upon a well-defined Patched Line Segment (PaLiS) representation for road graphs that holds geometric significance. To improve the accuracy and integrity of road network structures, various methods have been developed to refine road topology. They form multi-task branches by integrating CNNs with graph-based approaches. This integration ensures the preservation of contextual semantic information while maintaining the connectivity of roads. For instance, within an encoder–decoder framework, Li et al. [40] design a multi-task architecture that can concurrently predict the connectivity and segmentation of road maps. Zhang et al. [41] propose to extract road nodes and infer the connectivity between them. By learning the confidence map and the connectivity map for the nodes, the model predicts road nodes through a multi-task framework. With a GCN-based dual-view perceptor at its core, RDPGNet [42] integrates a CNN and a GCN to promote the exchange of multi-scale information.

Graph-based models, despite their potential, present notable drawbacks in road extraction. Firstly, graph construction is intricate. The precise definition of nodes, edges and their associated attributes demands profound domain knowledge, and inaccuracies in this stage can cascade, degrading the overall extraction quality. Secondly, graph-based models possess high computational complexity as well. Operations like graph traversal and message passing algorithms involve substantial computations, which not only slow down the extraction process but also call for powerful computational resources, restricting their application in real-time and large-scale scenarios.

2.3. ConvNets Based Architectures

With the advancement of convolutional networks (ConvNets) in deep learning, architectures such as U-Net [6,43], LinkNet [19,44], ResNet [25,45] and DeepLab [46,47] have been effectively applied in road extraction. There are several highly regarded encoder–decoder structures based on ConvNets for learning thin curvilinear road structures [43–56]. D-LinkNet [19] incorporates an encoder–decoder structure, dilated convolution and a pre-trained encoder to predict the road segmentation maps. The network is built upon the LinkNet [44] architecture, with the addition of dilated convolution layers in its central part. To achieve more precise segmentation results, Richer Unet [48] proposed a road detection approach by employing an enhanced detail recovery structure that can retain more abundant features. A Cascaded Attention DenseUNet (CADUNet) [49] used the core attention module to extract road areas, including the occluded parts, and used the global

attention module to enhance global contextual information about the road network. To solve the dense road occlusion problem, OARENet [50] is proposed to explicitly model the texture feature for road regions with dense occlusions. Ai et al. [51] applied variance and the coefficient of variation to the squeeze-and-excitation (SE) mechanism, designing a multi-parameter-guided SE module named MPGSE, which was then integrated into the D-LinkNet architecture. Hu et al. [52] developed the location-guided network (LGNet), inspired by strategies used in lane detection. The primary goal of LGNet was to address the common issue of disjointed extraction results that are prevalent in segmentation techniques. Qiu et al. [53] proposed a multimodal framework that synthesized GPS trajectory data with remote sensing features, significantly amplifying the model's discriminative power for road representation. Zhu et al. [43] introduced GCBNet, incorporating a global context-aware (GCA) block into the network to capture the global contextual information of roads. RADANet [54] incorporated a road augmentation module (RAM) and a deformable attention module (DAM) to effectively capture multi-scale semantic information. The RAM is designed to capture the semantic shape information of roads through four strip convolutions, while the DAM combines the sparse sampling capability of deformable convolution with a spatial self-attention mechanism.

These encoder–decoder-based methods also bring some drawbacks: on the one hand, these methods fuse multi-scale features by inserting well-designed but sophisticated modules into the networks, which introduce large parameters and limit their application in real world scenes. On the other hand, the downsampling operations of these networks on shallow feature maps inevitably lead to the loss of the crucial, detailed information of the thin, curvilinear roads, while upsampling operations on deep feature maps may cause the introduction of inaccurate semantic information. Hence, these methods fail to realize the effective aggregation of multi-scale features, and the generated fused feature cannot provide accurate road structural features for the decoder.

3. Proposed Method

In this section, we present the details of our Cascaded Efficient Road Network and introduce how we enhance the feature representation ability of our proposed model.

3.1. Overview

In this paper, we adopt the encoder–decoder structure as the overall framework. The main idea is to design a simple network architecture that can learn the importance maps of features at multiple levels and accurately aggregate them from a spatial perspective. Conventional hierarchical encoder–decoder architectures directly utilize the convolutional layer to reduce the resolution through downsampling operations and increase the channel number by a factor of two at each stage [26,45]. Since the downsampling operation enlarges the receptive field by sacrificing the detailed information of roads, we propose to down-sample the satellite image by a factor of 2 only twice, instead of four times. Apart from this, to address the issue of the shrinking receptive field caused by the reduced downsample operations, we replace the Convblocks with smoothed dilated convolution residual blocks. An attention-guided feature fusion sub-module is inserted as the bottleneck to fuse the multi-scale features to further strengthen the feature representation of roads. Finally, a simple decoder, which is made up of three convolutional layers, together with the aforementioned elements, form our encoder–decoder architecture, which we call Effi-RoadNet. Effi-RoadNet is cascaded twice to refine the extracted coarse road map and enhance the feature representation ability, as shown in Figure 2. In general, the proposed method can parallel patch-wise inference for large areas due to its low complexity. We demonstrate that

the sequential application of a cascade of the proposed Effi-RoadNet is able to predict road locations efficiently and effectively.

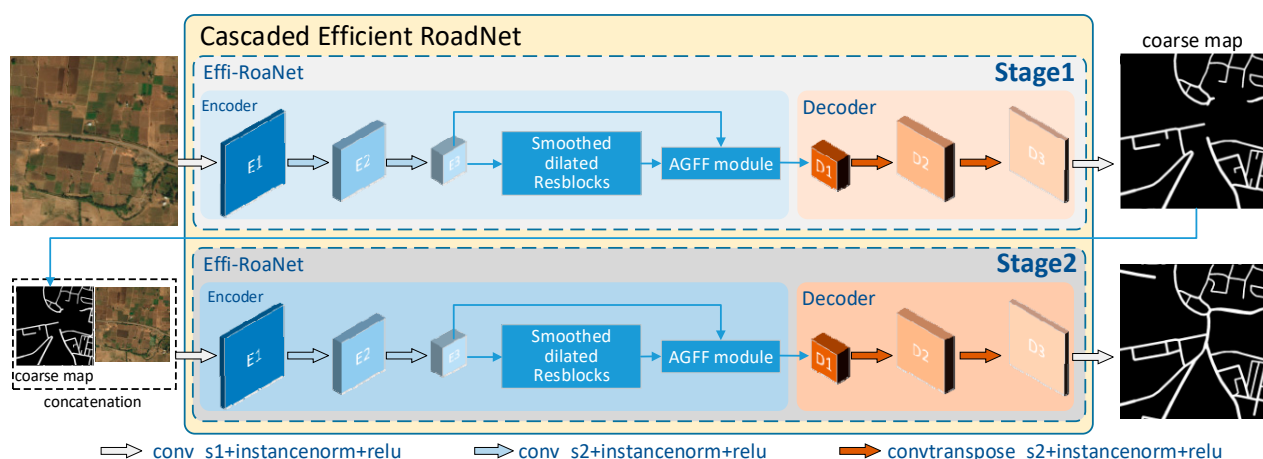


Figure 2. The overall framework of the proposed method.

3.2. Smoothed Dilated Residual Block

Given the narrowness, connectivity, complexity and extensive length of roads in remote sensing images, it is crucial to enhance the receptive field of the feature points in the central part of the network while preserving detailed information. Dilated convolution is a valuable technique to expand the receptive fields of feature points without reducing the resolution of feature maps. Furthermore, smoothed dilated convolution [33] addresses the gridding artifacts inherent in standard atrous convolutions by decomposing and smoothing the dilated convolution operations, thereby enhancing local consistency and reducing discontinuous receptive fields. Unlike conventional methods that rely on cascaded layers or hybrid dilation rates to mitigate gridding issues, this approach directly modifies the dilated convolution itself, achieving significant performance improvements with negligible additional parameters. As for road extraction tasks, smoothed dilated convolution is able to capture long-range dependencies in the image data while maintaining spatial resolution, enabling it to accurately delineate the road boundaries even in complex scenarios with various background textures and occlusions. The smoothing mechanism also helps in handling noise and irregularities in the remote sensing data, resulting in more reliable and detailed road extraction results compared to those of the standard atrous convolution. It should be noted that a separable and shared convolution is employed as the additional convolutional layer to integrate the neighboring information for each unit in the input feature maps. As shown in Figure 3, we stack the smoothed dilated convolution layer twice and add a shortcut connection to build a smoothed dilated residual block.

3.3. Attention-Guided Feature Fusion Sub-Module

The attention mechanism has been extensively utilized across a range of computer vision tasks. In this paper, we introduce a novel approach to fuse features from different levels by incorporating an additional attention-guided feature fusion (AGFF) module. Although local evidence is crucial for the identification of long-span, curvilinear, road-like features, a final estimate of the road necessitates a consistent comprehension of the scenes around roads. The orientation of the road, the layout of the surrounding buildings and tree shadows as well as the relationships of the adjacent joints are some of the numerous cues that are most effectively recognized at multiple scales within the image. The attention mechanism of AGFF has the capacity to aggregate contextual information from diverse levels. Consequently, it enhances the model's ability to learn contextual features.

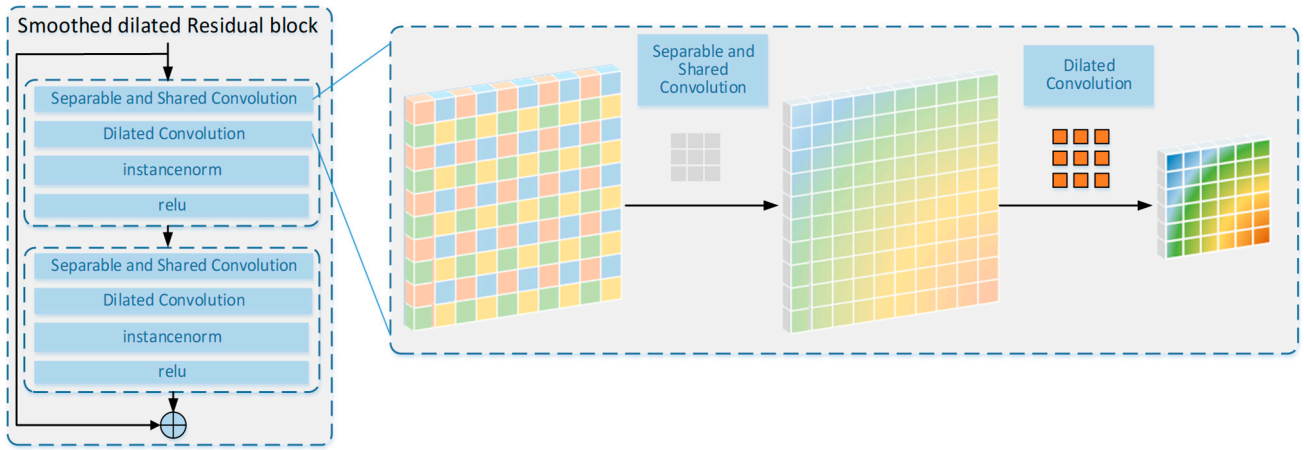


Figure 3. Smoothed dilated convolution residual block.

By analyzing the positional correlation between feature maps at different levels, AGFF directly computes a 3D attention map to accurately aggregate features from channel and spatial perspectives. A single convolution operation is employed to compress the channel-spatial information of different features. First, we extract multi-layer features from the encoder, which are denoted as F_l , F_m and F_h . Then, we feed them into the AGFF sub-module. More precisely, F_l represents the feature map that is fed into the first smoothed dilated convolution layer, F_m is the output generated by the third smoothed dilated convolution layer and F_h is the final output of the smoothed dilated convolution layers. After that, F_l , F_m and F_h are concatenated and fed into the feature fusion module. The output of the AGFF module consists of spatial attention maps (W_l , W_m , W_h), which correspond to each feature level, respectively. These maps are generated to selectively emphasize the importance maps of features at different scales. In the end, these three feature maps from different levels, namely F_l , F_m and F_h , are linearly combined with the regressed spatial attention maps. We describe the detailed operation below.

$$(W_l, W_m, W_h) = \sigma(\text{concat}(F_l, F_m, F_h)) \quad (1)$$

$$F_{\text{embed}} = W_l \otimes F_l + W_m \otimes F_m + W_h \otimes F_h \quad (2)$$

σ represents a straightforward convolutional layer with a kernel size of 3×3 . The input fed into σ is the concatenation of the feature maps F_l , F_m and F_h . $\otimes$ implies element-wise multiplication. Element-wise multiplication and addition are computationally efficient for fusing features from different levels. The combined feature map F_{embed} will then be fed into the decoder to produce the road map, as shown in Figure 4.

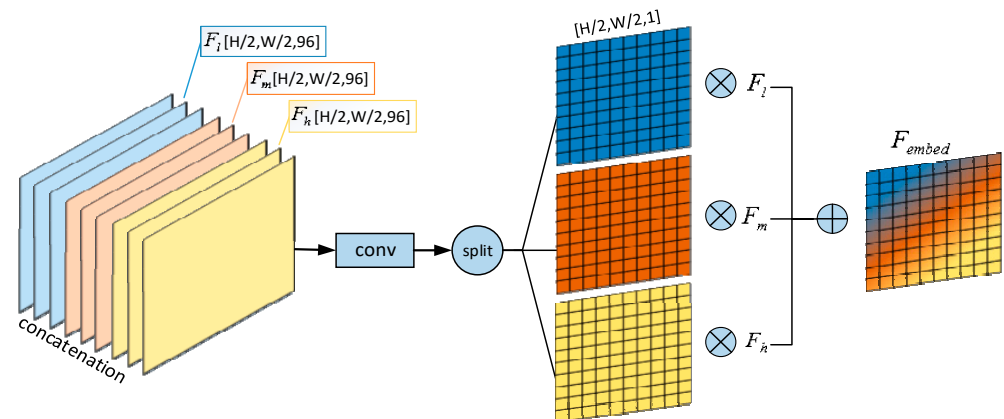


Figure 4. Illustration of attention-guided feature fusion sub-module.

3.4. *Effi-RoadNet*

With the development of AI chips, the bottleneck of the neural network's inference speed will no longer be the FLOPs or the number of parameters, as modern GPUs can easily perform parallel calculations with powerful computing capabilities. In contrast, the complex designs and extensive depths of these networks impede their inference speed. To address this issue, we introduce *Effi-RoadNet*, the architecture of which is depicted in Figure 2. We adopt the conventional autoencoder architecture, which consists of an encoder, a bottleneck and a decoder. Unlike conventional deep networks, our approach utilizes only one single layer at each stage, resulting in an extremely streamlined network architecture with a minimal number of layers. Apart from this, *Effi-RoadNet* eschews excessive depth, skip connections and complex operations such as self-attention. This results in a series of simplified network architectures that effectively mitigate inherent computation complexity.

To be more specific, we start by using three convolutional layers to encode the input image into feature maps. Notably, only the last two of these layers perform downsampling with stride 2. Symmetrically, in the decoder part, two deconvolutional layers with a stride of 2 are employed to upsample the feature map to its original resolution. This operation not only extracts the spatial context but also projects the fused feature map back into the image space. Once the network reaches the output resolution, two consecutive convolutional layers with 3×3 and 1×1 kernel sizes are utilized to generate the final target road map. Since only two downsample layers are incorporated in the encoding phase, the detailed information of the roads is well preserved. As a result, no more high-resolution information needs to be supplemented. We remove the skip connections between the encoder and the decoder because such operations would consume significant off-chip memory when they merge multi-level features.

We use six smoothed dilated residual blocks as the bottleneck, and each is made up of two smoothed dilated convolutional layers with a skip connection. The dilation rates are set as (2, 2, 2, 4, 4, 4, 1), respectively. The dilation rate of the last smoothed dilated residual block is 1, so that it is degraded to a normal residual block. The AGFF sub-module is inserted after the bottleneck to accurately aggregate context features from different levels. To balance performance and runtime efficiency, we configure the channel numbers of the intermediate convolutional layers to be 48, 64 and 96, respectively. With a simple yet minimal design, the proposed encoder–decoder architecture has the ability to aggregate different levels of context, enabling the fusion of features and the generation of pixel-wise predictions. It can fuse features and combine them to output pixel-wise predictions.

3.5. *Cascaded Effi-RoadNet*

We introduce a modification of the proposed *Effi-RoadNet* architecture, which we refer to as *CE-RoadNet* based on the idea behind Went [57]. Our overall architecture is constructed by stacking two *Effi-RoadNets* in an end-to-end configuration, where the output of the first *Effi-RoadNet* serves as a partial input to the second. This configuration enables the network to perform repeated encoder–decoder inference, allowing for the refinement and re-evaluation of initial estimates. The key point of this approach lies in the prediction of a coarse road map, upon which a loss function can be applied. The first road map is generated after passing through the first *Effi-RoadNet*, which converts images into coarse road maps. The second *Effi-RoadNet* enables the reprocessing of these high-level features (coarse maps) to further evaluate and reassess the higher order spatial relationships between the road and its surrounding scenes. Specifically, the pre-calculated coarse road map and the input remote sensing image are concatenated along the channel dimension to form the input for the second *Effi-RoadNet*. The first *Effi-RoadNet* is designed for coarse road detection tasks, leveraging its strong representational capabilities to effectively handle

complex backgrounds and occlusions caused by trees and cars. To refine the road topology and fix wrong predictions, the other Effi-RoadNet is cascaded to the former, fully utilizing the previously produced feature maps.

As demonstrated in Equation (3), for an input image x , the output obtained from passing it through the first Effi-RoadNet $\varphi^1(x)$ is concatenated with x , and this combined input is then passed through the second Effi-RoadNet $\varphi^2(x)$. This process can be mathematically represented as:

$$\phi(x) = \varphi^2(\text{concat}(x, \varphi^1(x))) \quad (3)$$

In practice, $\varphi^1(x)$ produces a rough prediction of the road's location. This prediction can then be employed by φ^2 as a form of attention map, enabling it to concentrate more intently on the areas of interest in the image, as illustrated in Figure 2. The cascade design incorporates twice the number of learnable parameters compared to the raw Effi-RoadNet. However, since the Effi-RoadNet involved in its definition contains only 1.39 M parameters each, the CE-RoadNet considered in this paper will have around 2.78 M parameters in total. This number is still less than that of the simplest architecture proposed to date for road segmentation, and is one order of magnitude lighter than the current state-of-the-art models.

4. Experimental Results

4.1. Datasets

The DeepGlobe dataset [29] is a comprehensive and widely used dataset in the field of geospatial image analysis and deep learning. It encompasses a diverse range of geographical regions around the world, providing a rich variety of landscapes, including urban areas, rural landscapes, forests and water bodies. There are 6226 images in it, each with a resolution of 1024×1024 . It is observed that certain portions of the images within the dataset are partially occluded. These images may have a severe detrimental impact on the performance of the models. Consequently, these occluded images are removed. We crop the rest of the images into 256×256 patches and obtain 56,192 images for training and 1172 images for testing.

The Massachusetts Road dataset [58] is another benchmark dataset used for road segmentation tasks. It is known for its high-quality annotations and is often used to evaluate the robustness of road extraction models across different geographical and environmental conditions. Following the same approach, we generate 25,619, 439 and 1380 images for the training, testing and validation sets, respectively.

4.2. Training Strategy

In this paper, we adopt the sum of binary cross entropy (BCE) and dice coefficient loss as the loss function. It is important to note that in the cascade case, the loss computed for the output of the second network is linearly combined with an auxiliary loss that is calculated for the output of the first network:

$$\ell(\phi(x), y) = \ell(\varphi^1(x), y) + \ell(\varphi^2(x), y) \quad (4)$$

The loss is back propagated and minimized using the AdamW optimization technique.

4.3. Experimental Settings

All the experiments are performed on two NVIDIA Tesla V100-PCIE GPUs (with 32 GB GPU memory) (NVIDIA, Santa Clara, CA, USA), with a fixed batch size of 16 samples. We set the initial learning rate to 6×10^{-3} , and when the loss does not decrease for five

consecutive epochs, we reduce it to one-fifth of its current value. The early stop strategy is used to avoid overfitting.

An augmentation strategy that includes horizontal/vertical flipping and shift scale rotation is used during the training phase. During the prediction phase, no test time augmentation (TTA) is employed. A threshold value of 0.5 is applied to generate the final output. In road segmentation tasks, 0.5 is often used as the threshold because it is intuitive for probability interpretation, offers symmetry in binary classification, is an empirical default and simplifies calculations.

4.4. Comparing Algorithms and Evaluation Metrics

To comprehensively verify the performance of the proposed method, a comparative analysis is conducted between the proposed method and five state-of-the-art road extraction algorithms, along with a retinal vessel segmentation method. The essential information of these methods is summarized as follows:

1. CasNet extracts consistent road areas and smooth road centerlines from very high resolution (VHR) remote sensing images simultaneously through a cascaded end-to-end convolutional neural network.
2. Dlink-net employs an encoder–decoder architecture, incorporating dilated convolutions and a pre-trained encoder to effectively perform the road extraction task.
3. SGCN integrates a split depth-wise separable graph convolutional network within an encoder–decoder architecture to capture better global contextual information.
4. CoANet proposes a connectivity attention network designed for road extraction from satellite imagery. It incorporates a strip convolution module to capture long-range context information from different directions, thereby aligning with the elongated and narrow shape of roads.
5. RCFS uses an improved encoder–decoder architecture, in which a multi-scale context extraction module is designed to boost the network’s inference ability by adding road context. In addition, multiple full-stage feature fusion modules are added to skip connections for accurate road structure information.
6. W-Net is a minimalistic model that achieves state-of-the-art performance in retinal vessel segmentation with significantly fewer parameters compared to other models.

It should be noted that the same learning rate scheduler and batch size are adopted for all the models.

All the approaches are evaluated with the TOPO [59], cIDice [60] and average path length similarity (APLS) [30]. TOPO randomly selects candidate vertices from the ground truth and identifies their corresponding elements in the prediction. It conducts a rigorous evaluation of the detailed topological correctness. This is achieved by comparing the similarity of the reachable subgraphs originating from the same vertex of the two graphs, with respect to precision, recall and F1 score. It is necessary to note that connectivity is an important factor that has to be considered for road extraction tasks, and APLS and cIDice are also adopted as evaluation metrics. APLS evaluates the quality of the predicted graph by measuring the similarity of the shortest paths between corresponding locations in the predicted and ground truth graphs. The CIDice loss function is designed to improve the connectivity of the results in road segmentation tasks by leveraging topological information. It calculates the similarity between the segmentation mask and its morphological skeleton, ensuring that the topological structure of the segmented object is preserved. To evaluate

the connectivity of the extracted road maps, we also use the *clDice* score, which is based on morphology. The following equations introduce the expression of the *clDice* coefficient:

$$T_p(S_p, V_L) = \frac{|S_p \cap V_L|}{|S_p|} \quad (5)$$

$$T_s(S_L, V_P) = \frac{|S_L \cap V_P|}{|S_L|} \quad (6)$$

$$clDice(V_P, V_L) = 2 \frac{T_p(S_p, V_L) T_s(S_L, V_P)}{T_p(S_p, V_L) + T_s(S_L, V_P)} \quad (7)$$

In this context, V_P denotes the predicted segmentation mask, and V_L stands for the ground truth mask. The skeletons S_P and S_L are extracted from V_P and V_L , respectively. Specifically, $T_p(S_p, V_L)$ indicates the topology precision, and $T_s(S_L, V_P)$ represents the relevant sensitivity.

APLS is defined as follows:

$$APLS = 1 - \frac{1}{P} \sum \min(1, \frac{|L(a, b) - L(\bar{a}, \bar{b})|}{L(a, b)}) \quad (8)$$

In this context, G represents the ground truth graph, while $\bar{G}$ stands for the predicted graph. P refers to the quantity of unique paths in G . $L(a, b)$ signifies the length of the path connecting the nodes a and b . Additionally, $\bar{a}$ and $\bar{b}$ are the nodes in $\bar{G}$ that are closest in location to nodes a and b . It should be noted that the value of APLS ranges from 0 (poor) to 1 (perfect).

4.5. Results and Comparative Analysis

Figures 5 and 6 depict the output maps of the proposed and existing state-of-the-art models on the DeepGlobe and Massachusetts Road datasets. As visualized, compared with other techniques, CE-RoadNet stands out for its outstanding adaptability to diverse scenarios. It provides accurate edge information of the roads, proving its superiority for both datasets. Such an improvement demonstrates that our CE-RoadNet is capable of extracting road features with remarkable invariance across diverse scenes, such as roads obscured by shadows, narrow roads, complex intersections or roads with obstructions. The arrows highlight especially challenging areas.

In Figure 5, for example, several roads are mislabeled in the ground truth road map where all methods predict part of the missing roads, among which CE-RoadNet achieves the best performance in preserving the continuity of the road topology (yellow arrows). Moreover, the predicted road masks by other methods suffer from varying degrees of fragmentation, while our CE-RoadNet exhibits better road connectivity (red arrows). CasNet encounters difficulties in accurately extracting roads in complex scenes with large areas of occlusion and interference; as shown in line 4, it cannot separate the road textures from those of the cluttered surroundings, leading to inaccurate segmentation results. In line 5, SGCN, CoANet and WNet face limitations in terms of recognizing the road-like features as roads. Among the seven models, the extraction results of WNet are the worst, since it is only a simple extension of UNet, which has a limited feature extraction ability. The degree of fragmentation of the extraction results and the number of extraction failures are the largest among all images, which is also consistent with the results of the quantitative analysis in Tables 1 and 2.

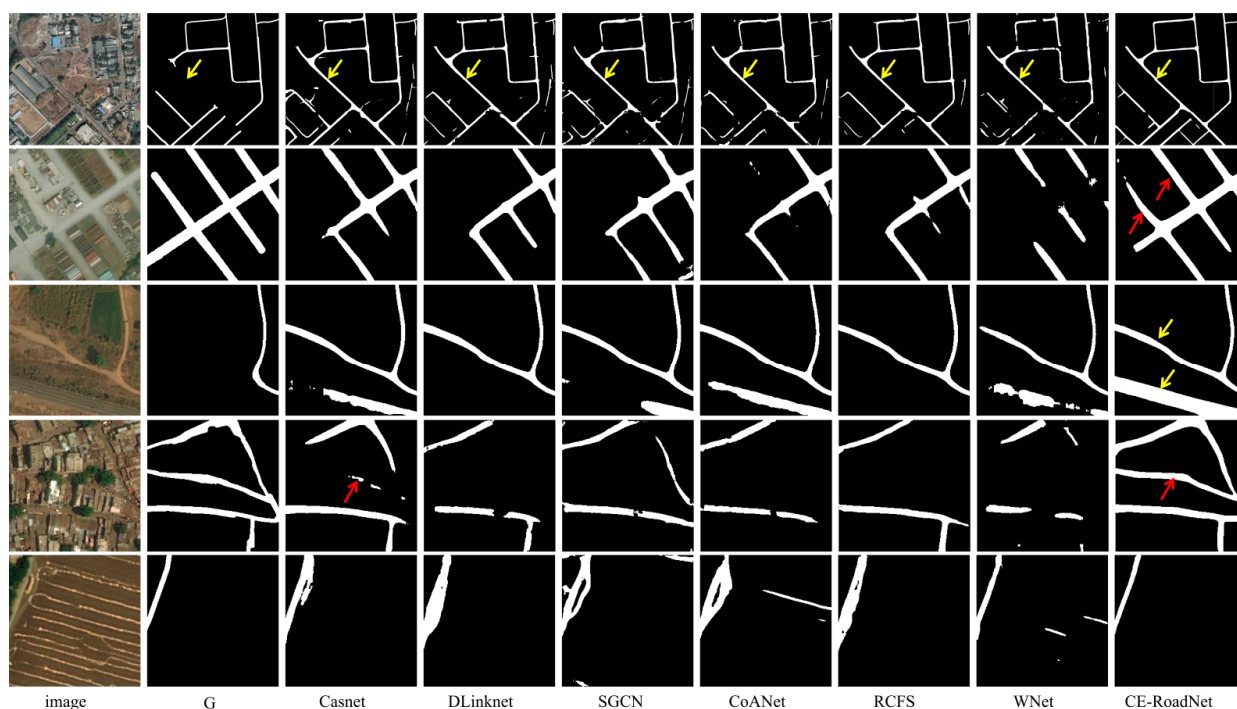


Figure 5. Qualitative evaluations between CE-RoadNet and comparative methods on the DeepGlobe road dataset. Maps in the first line are 1024×1024 while the rest are 256×256 .

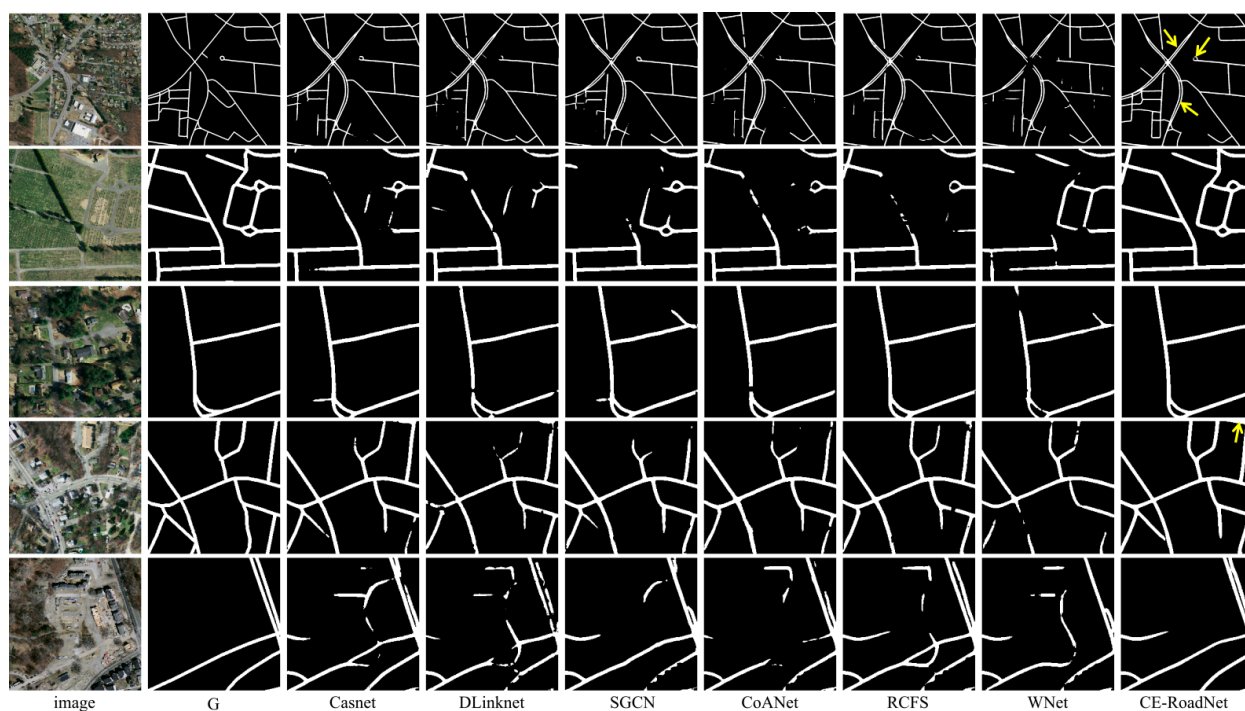


Figure 6. Qualitative evaluations between CE-RoadNet and comparative methods on the Massachusetts Road dataset. Maps in the first line are 768×768 while the rest are 256×256 .

In Figure 6, some dual carriageways are labeled as being a single way (yellow arrows). All methods extract the parallel freeways correctly, while CE-RoadNet gives superior results in extracting detailed local information such as ring roads and roads obscured by the surrounding trees or the shadows. In line 2, D-LinkNet, CoANet and RCFS cannot detect the roads that are close in color to the background. D-LinkNet struggles with accurately delineating road boundaries in complex urban environments like the ring roads in line 2. In line 3, CasNet

shows robustness against the occlusions of trees shadows to some extent, but the predicted road map still contains numerous fragments. For more complicated scenarios like that in line 4, the result predicted by CE-RoadNet more closely approximates ground truth, while all other methods tend to identify the roads as background and generate fragmented road maps. In general, CE-RoadNet is capable of predicting extremely precise road networks even under challenging circumstances, including multi-lane highways, intersections in heavily populated urban zones, winding roads with irregular geometries and complex overpasses.

Table 1. Comparative quantitative evaluation of different methods on Massachusetts dataset.

Method	Prec.	Rec.	F1	clDice	APLS
Casnet [18]	0.740	<u>0.779</u>	0.759	0.851	0.697
D-Linknet [19]	0.775	0.727	0.751	0.849	0.636
SGCN [22]	0.747	0.758	0.752	0.852	0.655
CoANet [20]	0.755	0.739	0.747	0.840	0.659
RCFSNet [21]	<u>0.782</u>	0.755	<u>0.768</u>	<u>0.862</u>	<u>0.701</u>
WNet [57]	0.770	0.702	0.734	0.828	0.607
CE-RoadNet	0.818	0.785	0.801	0.901	0.797

Table 2. Comparative quantitative evaluation of different methods on DeepGlobe dataset.

Method	Prec.	Rec.	F1	clDice	APLS
Casnet	0.604	<u>0.767</u>	0.676	0.428	0.505
D-Linknet	0.697	0.696	0.696	0.538	0.557
SGCN	0.636	0.721	0.676	0.442	0.509
CoANet	0.700	0.624	0.660	0.567	0.499
RCFSNet	<u>0.762</u>	0.660	<u>0.708</u>	<u>0.600</u>	<u>0.602</u>
WNet	0.680	0.574	0.623	0.522	0.389
CE-RoadNet	0.813	0.812	0.812	0.753	0.752

To dig deeper into the learning behavior and obtain more in-depth knowledge, we perform a qualitative analysis in the feature space. The normalized attention maps of both Effi-RoadNet stages are visualized in Figure 7. Features from low/medium/high levels are, respectively, multiplied with the corresponding attention map to produce an attention-guided feature embedding, which is further fed into the decoder. As can be seen, the AGFF sub-module enhances competition across neurons by fusing features from multiple levels, among which the low/medium level features show more low-frequency information and redundancy across channels. The Effi-RoadNet of stage 1 outputs a coarse road map in which the missing road in the ground truth map (white arrow) is also correctly detected. The Effi-RoadNet of stage 2 refines the result of stage 1 and reduces some noise (red arrow), thus maintaining the road topology perfectly. Feature diversity is promoted through the second stage in which the road texture becomes more distinct.

Tables 1 and 2 provide a summary of the evaluation metrics achieved by the proposed CE-RoadNet tested on the Massachusetts and DeepGlobe Road datasets. The best values are presented in bold, while the second-best values are underlined. It is evident that the proposed CE-RoadNet has a significant advantage over the comparative methods across all metrics on both datasets. D-LinkNet proposes a bridge module composed of dilated convolutions between the encoder and decoder to expand the receptive field. While D-LinkNet performs well in capturing multi-scale features, it also struggles with accurately delineating road boundaries in complex urban environments, as its bottleneck only contains several dilated convolution layers that may not fully address the challenges of fine-grained spatial detail preservation, resulting in a low recall score. CasNet simultaneously copes with the road detection and centerline extraction tasks to obtain more consistent road detection result, resulting in the highest score on the Recall metric except for CE-RoadNet. Although it

has two cascaded UNet, the final road centerlines are heavily dependent on the segmentation of the first network. Therefore, the feature extraction ability of this simple network architecture is very limited, making it difficult to maintain precise road connectivity in highly occluded or complex urban areas. SGCNNet presents a graph convolutional network in a CNN architecture to capture global contextual road information in channel and spatial features, but it has limitations in handling extremely complex road networks with dense occlusions or highly irregular road patterns, as its reliance on graph convolution and separable convolutions may struggle to fully capture fine-grained details in such scenarios. When dealing with roads with discontinuous segments in high-resolution imagery, CoANet shows degraded performance in road extraction tasks. The network's architecture is not robust enough to adapt to these complex scenarios, thus affecting the recall and F1 score. Meanwhile, the RCFSNet achieves the second-best performance in terms of precision, F1 cDice and APLS, which reflects the importance of extracting the long-distance road context. Despite leveraging contextual and multi-stage features, RCFSNet struggles to precisely extract roads in scenarios with a high level of textural complexity in the background, as the full-stage feature fusion may introduce noise to the decoder, leading to a low score on the Recall metric. WNet fails to capture the road context since multi-level features are not incorporated in the decoding procedure, leading to the inferior accuracy of its generated road results on both the TOPO and APLS scores. The proposed CE-RoadNet outperforms the second-best-listed RCFSNet method by 4.6%, 4% and 4.3% on the Massachusetts dataset and 6.3%, 2.3% and 1.5% on the DeepGlobe dataset in terms of the TOPO metric. Moreover, the proposed CE-RoadNet achieves the best APLS and cDice by incorporating the AGFF module and the cascade architecture together to extract local details and different levels of contextual information and aggregates them adaptively. CE-RoadNet is superior for road extraction tasks because it effectively combines multi-scale feature learning, contextual understanding and precise spatial detail preservation. These characteristics are essential for accurately extracting road networks from complex and varied environments, making it a more suitable choice than simply increasing the depth of CNN or adding modules with complex designs. Furthermore, the architecture of CE-RoadNet allows for the efficient fusion of features from multiple levels, combining low-level details with high-level semantics. This fusion is beneficial for road extraction, where both detailed edges and semantic understanding of the road structure are included.

The FLOPs, params and the inference latency of the models are also compared in Table 3 and Figure 8. The latency is measured on one NVIDIA Tesla V100-PCIE GPU (with 32 GB GPU memory) and Intel(R) Core(TM) i7-7700CPU (Intel, Santa Clara, CA, USA), with a fixed batch size of four samples. As shown, the sophisticated design of SGCN, CoANet and RCFS leads to increased computational complexity and memory usage, particularly when processing high-resolution images. CE-RoadNet endeavors to achieve the optimal balance between computational complexity and performance. It has the least parameters and only spends one-third of the floating-point operations than the most complex model, indicating that it is a far less complex model than most approaches. The diminished complexity of the model presented in this paper enhances its suitability for wide-range remote sensing image road detection scenarios.

Table 3. FLOPs and parameters of the models.

Method	FLOPs (G)	Params (M)	Latency (ms)		Input Size
			CPU	GPU	
Casnet	9.78	3.83	778	11.5	(4, 3, 256, 256)
D-Linknet	6.67	31.1	596	18.9	(4, 3, 256, 256)
SGCN	61.9	42.7	4734	80.3	(4, 3, 256, 256)
CoANet	69.3	59.1	4582	91.5	(4, 3, 256, 256)
RCFSNet	45.6	76.7	2539	45.4	(4, 3, 256, 256)
CE-RoadNet	13.08	2.78	1765	31.8	(4, 3, 256, 256)

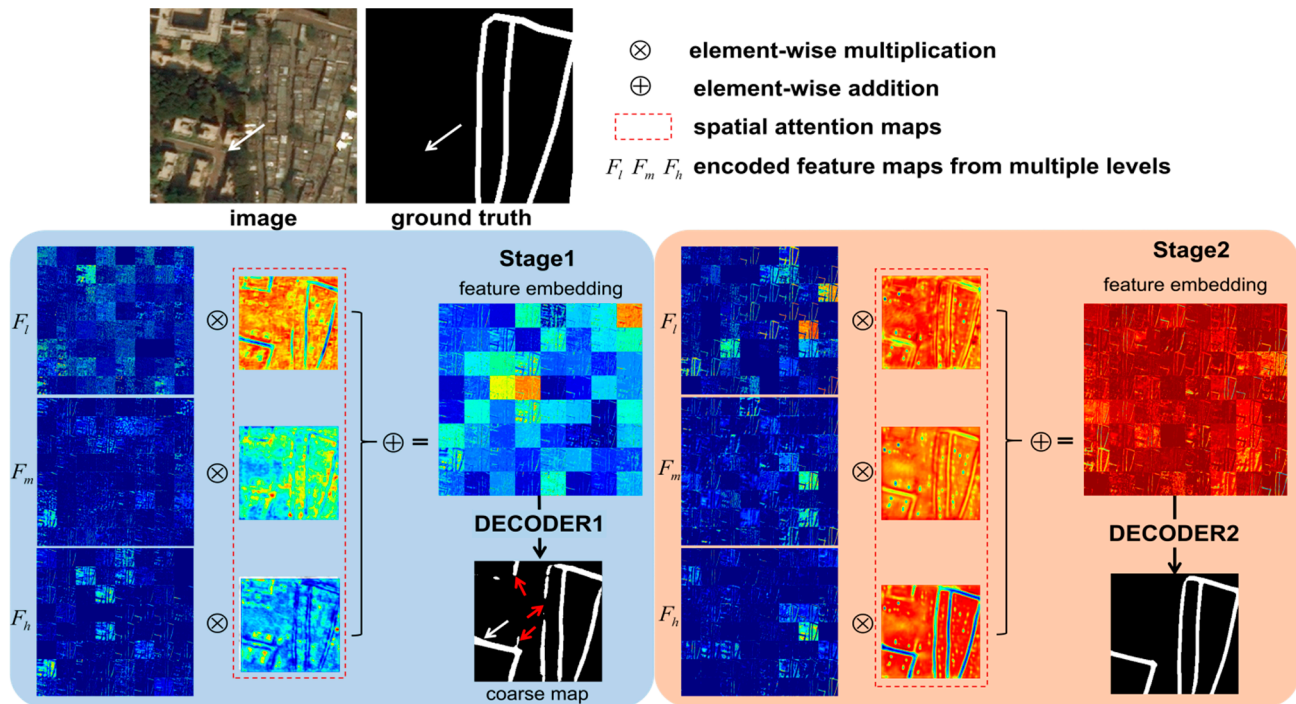


Figure 7. Feature activation visualization. Attention maps and the feature embeddings from stage 1 and stage 2. We visualize the normalized activation map for each feature channel in small squares. Notably, we display the first 64 channels in each visualization.

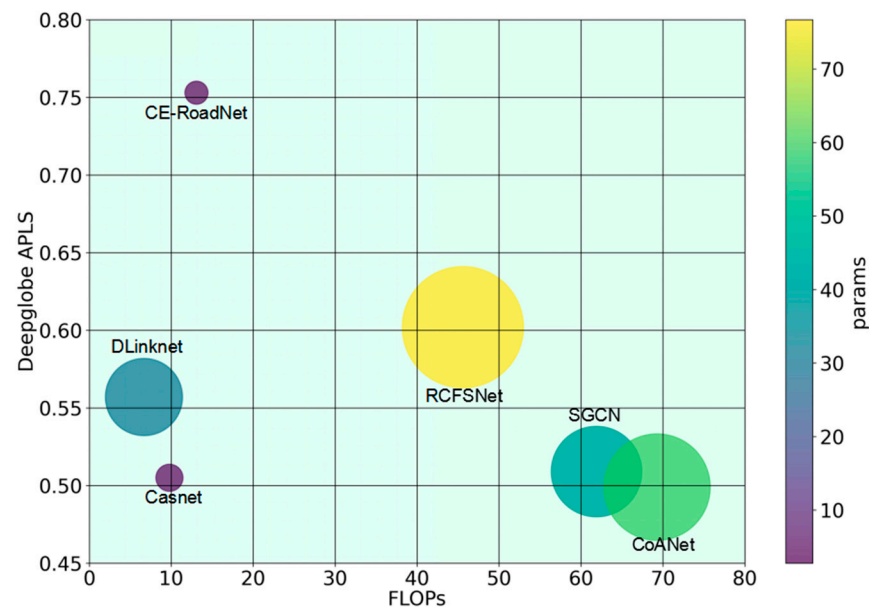


Figure 8. APLS results on the DeepGlobe dataset of the different models. Each bubble's area is proportional to the params of the corresponding model. All models here take 256^2 images as the input.

4.6. Ablation Experiments

In this section, we conduct further verification of the effectiveness of integrating the Effi-RoadNet, the cascade approach and skip connections in enhancing the performance of road extraction, as detailed in Table 4.

Effect of Effi-RoadNet: With the aim of further validating the generalizability of the Effi-RoadNet framework we put forward, we replace Effi-RoadNet with Unet and compare the performance on both datasets. Compared to Wnet, which is extended based on Unet, CE-RoadNet significantly improves performance across all metrics. This validates that

smoothed dilated convolution layers together with an AGFF sub-module can contribute to enhancing the capacity for aggregating information from a large receptive field.

Table 4. Comparative quantitative evaluation of baseline model under different configurations on the Massachusetts dataset.

Method	Prec.	Rec.	F1	clDice	APLS	FLOPs	Params
WNet	0.770	0.702	0.734	0.828	0.607	1.06	0.07
Effi-RoadNet	0.815	0.782	0.798	0.900	0.781	<u>6.52</u>	<u>1.39</u>
CE-RoadNet_cat	0.816	0.792	0.804	0.903	0.797	21.7	3.6
CE-RoadNet	<u>0.818</u>	0.785	<u>0.801</u>	0.901	0.797	13.08	2.78
CE-RoadNet_CBAM	0.820	0.733	0.78	0.88	0.751	13	2.77
CE-RoadNet3	0.817	<u>0.787</u>	<u>0.801</u>	<u>0.902</u>	<u>0.794</u>	19.63	4.17

Effect of cascade: CE-RoadNet is structured as a two-stage hierarchical encoder–decoder architecture. A single Effi-RoadNet spends only half the FLOPs and parameters than the baseline model. An acceptable decrease in the TOPO metrics occurs on the Massachusetts dataset when removing the cascade operation, since this refines the result of a single Effi-RoadNet and enhances the feature representation ability. With a significant increase of 2% in APLS, the cascade operation is thought to be more crucial in producing road maps with better connectivity. As a result, the cascade operation helps to predict difficult-to-recognize road areas by capturing accurate road contexts and yielding a significant improvement over a single Effi-RoadNet architecture. This result indicates that, except for the input image, pre-calculating the attention map of the road areas as the extra information is very beneficial for feature representation and network learning, leading to a predicted road map with better connectivity.

Effect of skip connection: To meet the requirement of efficiency, we remove the skip connection between the encoder and decoder. Here, we compare the performance under the two configurations. CE-RoadNet_cat indicates the model with skip connections to fuse the features of the same scale by concatenation and to alleviate the semantic gap of feature fusion. As is shown in Table 4, skip connections bring an increase of 66% and 29.5% in terms of FLOPs and parameters, but only lead to a negligible increase in Precision and clDice on the Massachusetts dataset. Since feature maps are downsampled by half only twice from the original image, the crucial details of the road are maintained to a great extent, so that not much detailed information needs to be supplemented during the decoding procedure. As a consequence, the increasing complexity of skip connections cannot bring considerable improvements in performance.

Effect of cascade times: We have respectively compared the network performance by stacking Effi-RoadNet once, twice and three times (denoted as Effi-RoadNet, CE-RoadNet and CE-RoadNet3, respectively). Undoubtedly, CE-RoadNet is the most concise and efficient, with the lowest computational complexity and parameter count, but CE-RoadNet3 has the best metric performance, despite having the highest computational complexity and parameter count. On the other hand, CE-RoadNet achieves a balance between efficiency and effectiveness; its precision, recall and F1 score are nearly on par with CE-RoadNet3, yet its computational complexity is only 66% that of CE-RoadNet3.

Effect of AGFF: To test the effectiveness of the proposed AGFF module, we replace it with CBAM [61], a simple yet effective attention module for feed-forward convolutional neural networks that can improve the representation power of CNN networks, and we denote it as CE-RoadNet_CBAM. The experimental results demonstrate that although AGFF experiences a slight decrease in Precision, it achieves improvements of 7% in recall and 6% in APLS, respectively. This indicates that AGFF has advantages in terms of comprehensive performance improvement compared with other attention mechanisms.

5. Conclusions

In this paper, we propose a novel cascaded efficient road extraction network (CE-RoadNet). Specifically, we introduce Efficient RoadNet (Effi-RoadNet), which adopts only two downsample layers with stride 2 in the encoder to minimize information loss and preserve more spatial details. To counteract the shrinking receptive field caused by reduced downsampling, we replace the conventional residual blocks with smoothed dilated convolution residual blocks, enhancing the model's ability to capture multi-scale features. Furthermore, we introduce an attention-guided feature fusion (AGFF) sub-module in the bottleneck, which dynamically aggregates road features across various locations and hierarchical levels, providing the decoder with precise and comprehensive visual features. CE-RoadNet is implemented in a two-stage cascaded architecture, where Effi-RoadNet is applied iteratively to refine the coarse road map generated in the first stage. This cascade architecture significantly improves the completeness of the predicted roads and effectively closes the gaps in the results. Compared to existing methods, CE-RoadNet generates more accurate and well-connected road maps, demonstrating superior performance in challenging scenarios. This study fills critical gaps in the existing research by introducing innovations such as the AGFF module for dynamic feature fusion, smoothed dilated convolution residual blocks for enhanced receptive fields and a cascaded architecture for iterative refinement. These contributions differentiate CE-RoadNet from previous approaches and advance the state-of-the-art model for road extraction. In future work, we aim to explore the integration of other state-of-the-art foundational models within the context of road segmentation tasks and challenge the road vectorization task.

Author Contributions: Conceptualization, K.-N.C., W.N., H.Z. and J.W.; methodology, K.-N.C.; validation, K.-N.C. and Z.Y.; investigation, K.-N.C.; writing—original draft preparation, K.-N.C.; writing—review and editing, K.-N.C., W.N., H.Z., J.W. and X.X.; supervision, K.-N.C., W.N., H.Z. and J.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The Deepglobe road dataset is available at <http://deepglobe.org/challenge.html>. The Massachusetts road dataset is available at <https://www.cs.toronto.edu/~vmnih/data/>.

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

References

1. He, Y.; Wang, H.; Zhang, B. Color-based road detection in urban traffic scenes. *IEEE Trans. Intell. Transp. Syst.* **2004**, *5*, 309–318.
2. Chai, D.; Forstner, W.; Lafarge, F. Recovering line-networks in images by junction-point processes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 1894–1901.
3. Laptev, I.; Mayer, H.; Lindeberg, T.; Eckstein, W.; Steger, C.; Baumgartner, A. Automatic extraction of roads from aerial images based on scale space and snakes. *Mach. Vis. Appl.* **2000**, *12*, 23–31. [[CrossRef](#)]
4. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [[CrossRef](#)] [[PubMed](#)]
5. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
6. 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; Proceedings, Part III 18; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
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. Zhang, Z.; Liu, Q.; Wang, Y. Road extraction by deep residual u-net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753. [[CrossRef](#)]
9. Hetang, C.; Xue, H.; Le, C.; Yue, T.; Wang, W.; He, Y. Segment anything model for road network graph extraction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 2556–2566.

10. Bastani, F.; He, S.; Abbar, S.; Alizadeh, M.; Balakrishnan, H.; Chawla, S.; Madden, S.; DeWitt, D. Roadtracer: Automatic extraction of road networks from aerial images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4720–4728.
11. 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.
12. Xu, Z.; Liu, Y.; Gan, L.; Sun, Y.; Wu, X.; Liu, M.; Wang, L. Rngdet: Road network graph detection by transformer in aerial images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4707612. [\[CrossRef\]](#)
13. Xu, Z.; Liu, Y.; Sun, Y.; Liu, M.; Wang, L. Rngdet++: Road network graph detection by transformer with instance segmentation and multi-scale features enhancement. *IEEE Robot. Autom. Lett.* **2023**, *8*, 2991–2998. [\[CrossRef\]](#)
14. Batra, A.; Singh, S.; Pang, G.; Basu, S.; Jawahar, C.; Paluri, M. Improved road connectivity by joint learning of orientation and segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10385–10393.
15. He, S.; Bastani, F.; Jagwani, S.; Alizadeh, M.; Balakrishnan, H.; Chawla, S.; Elsharif, M.M.; Madden, S.; Sadeghi, M.A. Sat2graph: Road graph extraction through graph-tensor encoding. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XXIV 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 51–67.
16. Tan, Y.Q.; Gao, S.H.; Li, X.Y.; Cheng, M.M.; Ren, B. Vecroad: Point-based iterative graph exploration for road graphs extraction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 8910–8918.
17. Zao, Y.; Zou, Z.; Shi, Z. Topology-guided road graph extraction from remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *62*, 4700614. [\[CrossRef\]](#)
18. Cheng, G.; Wang, Y.; Xu, S.; Wang, H.; Xiang, S.; Pan, C. Automatic road detection and centerline extraction via cascaded end-to-end convolutional neural network. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3322–3337. [\[CrossRef\]](#)
19. 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.
20. Mei, J.; Li, R.J.; Gao, W.; Cheng, M.M. CoANet: Connectivity attention network for road extraction from satellite imagery. *IEEE Trans. Image Process.* **2021**, *30*, 8540–8552. [\[CrossRef\]](#) [\[PubMed\]](#)
21. Yang, Z.; Zhou, D.; Yang, Y.; Zhang, J.; Chen, Z. Road extraction from satellite imagery by road context and full-stage feature. *IEEE Geosci. Remote Sens. Lett.* **2022**, *20*, 8000405. [\[CrossRef\]](#)
22. Zhou, G.; Chen, W.; Gui, Q.; Li, X.; Wang, L. Split depth-wise separable graph-convolution network for road extraction in complex environments from high-resolution remote-sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5614115. [\[CrossRef\]](#)
23. 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.
24. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
25. Yang, Z.; Zhang, W.; Li, Q.; Ni, W.; Wu, J.; Wang, Q. C²Net: Road Extraction via Context Perception and Cross Spatial-Scale Feature Interaction. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5647011. [\[CrossRef\]](#)
26. Woo, S.; Debnath, S.; Hu, R.; Chen, X.; Liu, Z.; Kweon, I.S.; Xie, S. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 16133–16142.
27. Guo, M.H.; Lu, C.Z.; Liu, Z.N.; Cheng, M.M.; Hu, S.M. Visual attention network. *Comput. Vis. Media* **2023**, *9*, 733–752. [\[CrossRef\]](#)
28. Yang, J.; Li, C.; Dai, X.; Gao, J. Focal modulation networks. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 4203–4217.
29. 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.
30. Van Etten, A.; Lindenbaum, D.; Bacastow, T.M. Spacenet: A remote sensing dataset and challenge series. *arXiv* **2018**, arXiv:1807.01232.
31. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part v 13; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
32. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [\[CrossRef\]](#)

33. Wang, Z.; Ji, S. Smoothed dilated convolutions for improved dense prediction. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018; pp. 2486–2495.
34. Ge, C.; Nie, Y.; Kong, F.; Xu, X. Improving road extraction for autonomous driving using swin transformer unet. In Proceedings of the 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC), Macau, China, 8–12 October 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1216–1221.
35. Yang, Z.; Zhou, D.; Yang, Y.; Zhang, J.; Chen, Z. TransRoadNet: A novel road extraction method for remote sensing images via combining high-level semantic feature and context. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6509505. [\[CrossRef\]](#)
36. Zhang, X.; Ma, X.; Yang, Z.; Liu, X.; Chen, Z. A context-aware road extraction method for remote sensing imagery based on transformer network. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 5511205. [\[CrossRef\]](#)
37. Yang, Z.X.; You, Z.H.; Chen, S.B.; Tang, J.; Luo, B. Semisupervised edge-aware road extraction via cross teaching between CNN and transformer. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 8353–8362. [\[CrossRef\]](#)
38. Jamali, A.; Roy, S.K.; Li, J.; Ghamisi, P. Neighborhood attention makes the encoder of resunet stronger for accurate road extraction. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 6003005. [\[CrossRef\]](#)
39. Xu, J.; Xu, B.; Xia, G.S.; Dong, L.; Xue, N. Patched line segment learning for vector road mapping. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 6288–6296.
40. Li, T.; Ye, S.; Li, R.; Fu, Y.; Yang, G.; Pan, Z. Topology-aware road extraction via multi-task learning for autonomous driving. In Proceedings of the 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC), Bilbao, Spain, 24–28 September 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 2275–2281.
41. Zhang, J.; Hu, X.; Wei, Y.; Zhang, L. Road topology extraction from satellite imagery by joint learning of nodes and their connectivity. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5602613. [\[CrossRef\]](#)
42. Liu, G.; Shan, Z.; Meng, Y.; Akbar, T.A.; Ye, S. RDPGNet: A road extraction network with dual-view information perception based on GCN. *J. King Saud Univ.-Comput. Inf. Sci.* **2024**, *36*, 102009. [\[CrossRef\]](#)
43. Zhu, Q.; Zhang, Y.; Wang, L.; Zhong, Y.; Guan, Q.; Lu, X.; Zhang, L.; Li, D. A global context-aware and batch-independent network for road extraction from VHR satellite imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *175*, 353–365. [\[CrossRef\]](#)
44. 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; IEEE: Piscataway, NJ, USA, 2017; pp. 1–4.
45. 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.
46. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
47. Lin, Y.; Xu, D.; Wang, N.; Shi, Z.; Chen, Q. Road extraction from very-high-resolution remote sensing images via a nested SE-Deeplab model. *Remote Sens.* **2020**, *12*, 2985. [\[CrossRef\]](#)
48. Zao, Y.; Shi, Z. Richer U-Net: Learning more details for road detection in remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 3003105. [\[CrossRef\]](#)
49. Li, J.; Liu, Y.; Zhang, Y.; Zhang, Y. Cascaded attention DenseUNet (CADUNet) for road extraction from very-high-resolution images. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 329. [\[CrossRef\]](#)
50. Yang, R.; Zhong, Y.; Liu, Y.; Lu, X.; Zhang, L. Occlusion-aware road extraction network for high-resolution remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5619316. [\[CrossRef\]](#)
51. Ai, J.; Hou, S.; Wu, M.; Chen, B.; Yan, H. MPGSE-D-LinkNet: Multiple-parameters-guided squeeze-and-excitation integrated D-LinkNet for road extraction in remote sensing imagery. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 5508205. [\[CrossRef\]](#)
52. Hu, J.; Gao, J.; Yuan, Y.; Chanussot, J.; Wang, Q. LGNet: Location-guided network for road extraction from satellite images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5619112. [\[CrossRef\]](#)
53. Qiu, Y.; Lin, C.; Mei, J.; Sun, Y.; Lu, H.; Xu, J. Lightweight cross-modal information measure and propagation for road extraction from remote sensing image and trajectory/LiDAR. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4504716. [\[CrossRef\]](#)
54. 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. [\[CrossRef\]](#)
55. Yang, Z.; Li, Q.; Yuan, Y.; Wang, Q. HCNet: Hierarchical feature aggregation and cross-modal feature alignment for remote sensing image captioning. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5624711. [\[CrossRef\]](#)
56. Wang, Q.; Yang, Z.; Ni, W.; Wu, J.; Li, Q. Semantic-Spatial Collaborative Perception Network for Remote Sensing Image Captioning. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5649912. [\[CrossRef\]](#)
57. Galdran, A.; Anjos, A.; Dolz, J.; Chakor, H.; Lombaert, H.; Ayed, I.B. The little w-net that could: State-of-the-art retinal vessel segmentation with minimalistic models. *arXiv* **2020**, arXiv:2009.01907.
58. Mnih, V. *Machine Learning for Aerial Image Labeling*; University of Toronto: Toronto, ON, Canada, 2013.

59. Biagioni, J.; Eriksson, J. Inferring road maps from global positioning system traces: Survey and comparative evaluation. *Transp. Res. Rec.* **2012**, 2291, 61–71. [[CrossRef](#)]
60. Shit, S.; Paetzold, J.C.; Sekuboyina, A.; Ezhov, I.; Unger, A.; Zhylka, A.; Pluim, J.P.; Bauer, U.; Menze, B.H. clDice-a novel topology-preserving loss function for tubular structure segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 16560–16569.
61. 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.

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.