

Article

PCCAU-Net: A Novel Road Extraction Method Based on Coord Convolution and a DCA Module

Xiaoqin Xue ¹, Chao Ren ^{1,2,*} , Anchao Yin ¹ , Ying Zhou ¹, Yuanyuan Liu ¹, Cong Ding ¹ and Jiakai Lu ¹

¹ College of Geomatics and Geoinformation, Guilin University of Technology, Guilin 541006, China; 2120222038@glut.edu.cn (X.X.); yinanchao@glut.edu.cn (A.Y.); zhouying01@glut.edu.cn (Y.Z.); liuyyuan@glut.edu.cn (Y.L.); 1020221944@glut.edu.cn (C.D.); 2120222009@glut.edu.cn (J.L.)

² Guangxi Key Laboratory of Spatial Information and Geomatics, Guilin 541106, China

* Correspondence: renchao@glut.edu.cn

Abstract: In the domain of remote sensing research, the extraction of roads from high-resolution imagery remains a formidable challenge. In this paper, we introduce an advanced architecture called PCCAU-Net, which integrates Pyramid Pathway Input, CoordConv convolution, and Dual-Input Cross Attention (DCA) modules for optimized performance. Initially, the Pyramid Pathway Input equips the model to identify features at multiple scales, markedly enhancing its ability to discriminate between roads and other background elements. Secondly, by adopting CoordConv convolutional layers, the model achieves heightened accuracy in road recognition and extraction against complex backdrops. Moreover, the DCA module serves dual purposes: it is employed at the encoder stage to efficiently consolidate feature maps across scales, thereby fortifying the model's road detection capabilities while mitigating false positives. In the skip connection stages, the DCA module further refines the continuity and accuracy of the features. Extensive empirical evaluation substantiates that PCCAU-Net significantly outperforms existing state-of-the-art techniques on multiple benchmarks, including precision, recall, and Intersection-over-Union(IoU). Consequently, PCCAU-Net not only represents a considerable advancement in road extraction research, but also demonstrates vast potential for broader applications, such as urban planning and traffic analytics.

Keywords: road extraction; pyramid pathway input; CoordConv; Dual-Input Cross Attention (DCA) Modules



Citation: Xue, X.; Ren, C.; Yin, A.; Zhou, Y.; Liu, Y.; Ding, C.; Lu, J.

PCCAU-Net: A Novel Road Extraction Method Based on Coord Convolution and a DCA Module. *Appl. Sci.* **2024**, *14*, 1634. <https://doi.org/10.3390/app14041634>

Academic Editors: Sungho Kim and Nektarios A. Valous

Received: 27 November 2023

Revised: 23 January 2024

Accepted: 15 February 2024

Published: 18 February 2024



Copyright: © 2024 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

1.1. Related Works

In the swiftly advancing field of remote sensing technology, we now have unprecedented access to high-resolution images of the Earth's surface, offering invaluable insights into geography [1], ecology [2] and urbanization [3]. Despite these advances, the processing of high-resolution remote sensing images presents myriad challenges, particularly in applications like road extraction, building detection, and land use classification [4,5]. Accurate, rapid, and automated road information extraction from these images is pivotal, not only for smart city construction, transportation planning, and autonomous driving technologies, but also plays a crucial role in emergency management [6], environmental monitoring [7], and military reconnaissance [8]. Traditional remote sensing image processing methods mainly use template matching approaches [9], knowledge-driven approaches, object-oriented approaches [10], and path morphology approaches [11]. The breakthroughs in computer vision via deep learning, especially convolutional neural networks (CNNs), have ushered in new opportunities for remote sensing image processing [12]. CNNs autonomously learn representative features from data, thus achieving impressive performance across diverse tasks. However, conventional CNN architectures are not entirely equipped to handle the unique challenges presented by high-resolution remote sensing images, such as large homogeneous areas, complex background structures, and intricate target features [13].

Subsequent innovations like Fully Convolutional Networks (FCNs) adapted CNNs for pixel-level segmentation by replacing fully connected layers with convolutional layers [14], but limitations such as a loss of detail due to continuous down-sampling remain unresolved [15]. U-Net [16] introduced skip connections for the first time. Its encoder reduces the spatial dimensionality and captures spatial detail information, while the decoder reuses the low-level features from the encoding stage to gradually restore the input size and recover spatial location information. Since then, people have continuously made improvements to the U-Net architecture to alleviate these challenges and solve the gradient problem through residual structures [17].

Hybrid networks like ResU-Net have further combined the advantages of U-Net and Res-Net, stabilizing gradients while facilitating information exchange through shortcut connections [18]. Techniques like dilated convolutions in DeepLab series networks have allowed for multi-scale feature learning while expanding the receptive field [19–22]. This method utilizes dilated convolutions instead of up-sampling and combining dilated convolutions with atrous spatial pyramid pooling modules to augment the receptive field and obtain multi-scale features and enhance the model's multi-scale prediction capabilities. However, the feature extraction method based on dilated convolutions can easily lose information related to small-scale targets as the convolution kernels solely extract features from restricted regions. Wang et al. [23] introduced the HRNet, which employs a multi-branch structure to simultaneously maintain high- and low-resolution features and repeatedly performed multi-scale fusion to generate rich high-resolution representations. Nevertheless, in occluded images, some resolutions may exhibit information deficits due to partial regions being obscured, ultimately affecting the effectiveness of the resulting feature maps. In parallel, attention mechanisms have been introduced in machine translation tasks [24]. Woo proposed the Convolutional Block Attention Module (CBAM) based on the attention mechanism in 2018; this is a lightweight module that pays attention to the importance of both space and channel [25]. Li obtained the contextual information of the road through global attention, used core attention to obtain multi-scale information, improved the connectivity of road extraction, and restored the information at the occlusion place. At the same time, he improved the cross-entropy function and proposed an adaptive loss function in order to effectively solve the problem of road and non-road areas in training samples [26]. Models like SENet [27] and CCNet [28] incorporate attention mechanisms capable of enhancing semantic segmentation accuracy. Novel attention-based networks, such as CSANet [29], capture long-range dependencies at a lower computational cost. Various other approaches, like HA-RoadFormer [30] and MSFANet [31], further enrich feature representations by leveraging long-range dependencies or multi-spectral data, thereby mitigating the impact of noise and redundant features. Alshaikhli [32] effectively obtained pixel values and optimized prediction results by combining residual blocks and attention. Qiqi Zhu [33] proposed a global context-aware and batch-independent network, using a global context-aware network embedded in the encoder–decoder structure to extract complete and continuous road networks. Shao Z [34] proposed a dual-task end-to-end convolutional neural network, using cumulative convolution and a pyramid scene-parsing pool module, expanding the network reception field, integrating multi-level features, and obtaining richer road information. Yu Rong [35] proposed a new context-enhanced and self-attention capsule feature pyramid network, which integrates context enhancement and self-attention modules, uses multi-scale context attributes and channel information feature enhancement, and enhances the robustness of feature representation. However, existing methods often fall short in adequately fusing multi-scale features [36] and in contextual modeling across encoding and decoding phases [37], thereby limiting their learning capability and attention optimization for road extraction tasks. The development of more effective algorithms for remote sensing image processing remains an area of active research.

1.2. Innovation Statement

To surmount the persistent challenges associated with road extraction in high-resolution remote sensing imagery, researchers have devised a plethora of deep learning methodologies tailored specifically to remote sensing applications. Among these, the structurally refined PCCAU-Net stands as a recent breakthrough, incorporating Pyramid Pathway Input, CoordConv convolution, and Dual Input Cross Attention (DCA) modules. This innovative amalgamation facilitates the more efficacious capture and integration of multi-scale features, thereby augmenting both the accuracy and robustness of road extraction.

Distinctive contributions and attributes of our model are as follows:

1. Multi-scale feature and spatial context fusion: PCCAU-Net employs Pyramid Pathway Input and CoordConv convolution, achieving a seamless assimilation of features from across multiple scales down to fine-grain levels, along with precocious spatial context recognition. This holistic design strategy assures precise road localization and extraction, even amidst complex settings.
2. At the critical encoder stage, the model incorporates a Dual-Input Cross Attention (DCA) module. This not only augments the nimble integration of feature maps across various scales but also sharpens the model's focus on road center detection, effectively attenuating false positives induced by obstructions, entanglements, and noise.
3. Across key performance metrics such as precision, recall, and Intersection-over-Union (IoU), PCCAU-Net demonstrates a clear supremacy over existing mainstream approaches, vindicating its potential applicability in the realm of remote sensing data processing.

This manuscript elucidates the design principles, empirical validations, and potential applications of this model, aiming to offer substantive insights for future endeavors in remote sensing data processing. The article is structured as follows: The introduction delineates the backdrop and challenges of road extraction in high-resolution remote sensing imagery and elucidates why traditional deep learning models may encounter difficulties. Subsequently, we present our solution, an innovative deep learning model named PCCAU-Net, endowed with Pyramid Pathway Input and Dual Input Cross Attention (DCA) modules. In the Methods section, we provide a comprehensive description of the PCCAU-Net architecture and its operational mechanics, including its cornerstone components like CoordConv and DCA modules. In the Results section, we introduce the two road datasets selected for the experiment, Massachusetts Roads Dataset and Deep-Globe Road Dataset, which provide rich image data for road-related computer vision tasks. Secondly, through comparative analysis and ablation experiments, various performance indicators are compared to demonstrate the performance of our model on these two primary remote sensing datasets, confirming its effectiveness and superiority. The Discussion section provides a detailed review of the experimental results and analyzes the advantages and potential limitations of PCCAU-Net. Finally, the conclusion summarizes the key contributions and discoveries of this study, pointing to directions for future research and potential improvements.

2. Methods

In the realm of high-resolution remote sensing, the task of road extraction persistently presents a challenging conundrum. To mitigate this, we elaborate upon a novel and architecturally refined deep learning model, designated as PCCAU-Net. This model amalgamates a variety of cutting-edge techniques with the intent of capturing the intricate spatial contextual relationships inherent in high-resolution remote sensing imagery, thereby augmenting the accuracy of road extraction.

2.1. Overview of PCCAU-Net

Designed explicitly to address the intrinsic complexities of high-resolution remote sensing imagery, PCCAU-Net is a deep learning model that features a unique multi-scale, multi-tiered strategy. Traditional deep learning approaches often encounter difficulties when confronted with diverse terrains, such as urban and rural landscapes [38]. In contrast,

PCCAU-Net excels in road extraction tasks, executing them with both greater precision and robustness. One of the model's salient features is its Pyramid Pathway Input. This component not only processes imagery at its original resolution but also automatically rescales it across four distinct levels, creating an input feature pyramid. This multi-scale input mechanism enables the model to capture information comprehensively, from micro- to macroscales, meeting the demands for feature extraction in complex landscapes and road structures encountered in high-resolution remote sensing imagery. To augment the model's geospatial perception further, CoordConv convolutional layers are initially applied to the five input images sourced from the pyramid pathway. This ensures an equitable treatment of geospatial information alongside traditional pixel values, affording the model deep insights into geographic positioning at early stages. Additionally, the DCA module furnishes unprecedented feature integration capabilities. At the encoder stage, it effectively amalgamates information from two different pyramid pathways, which is crucial for handling common issues like occlusion, entanglement, and noise in high-resolution imagery. Further, at the decoder stage, the DCA serves as a skip connection, facilitating feature integration between the decoder and its corresponding encoder, and thereby ensuring the full recovery of image details.

In summary, PCCAU-Net not only surmounts the limitations inherent to traditional methods but also heralds a new era in the processing of high-resolution remote sensing imagery, offering an innovative and exceptionally effective solution. The PCCAU-Net network structure is shown in Figure 1.

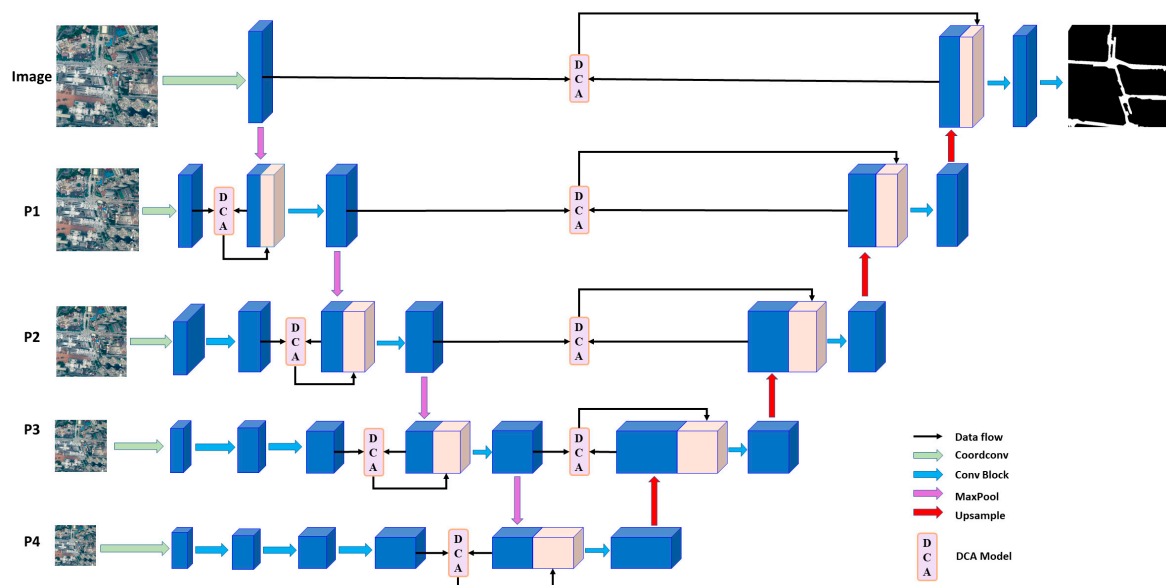


Figure 1. PCCAU-Net network structure diagram, in which Image represents the original size image input, P1, P2, P3 and P4 represent four different scales of pyramid path input, and the combination of blue and light pink feature blocks represents the Concat connection operation.

2.2. Pyramid Path Input

While the traditional U-Net architecture is undeniably robust, it often compromises spatial granularity through its encoding stages, composed of serial convolutions, and pooling layers in order to facilitate abstract semantic features [16]. Although beneficial for certain applications, this approach may inadvertently obfuscate or diminish critical spatial characteristics essential for precise analyses of high-resolution remote sensing data. Our remedy is the incorporation of a Pyramid Input Pathway within the PCCAU-Net framework. In this innovative configuration, the encoder of PCCAU-Net is not restricted to merely processing the raw image input. Instead, it fabricates four distinct pyramid levels (P1, P2, P3, and P4) by directly scaling the original image to the necessary respective sizes.

Each level captures data at a unique spatial resolution, ensuring the network consistently acquires features spanning both broad context and granular detail. This multi-resolution strategy furnishes the model with a comprehensive field of vision, empowering it to judiciously balance macro- and micro-level information during the deep learning phase. The outcome is twofold: it augments the model's contextual understanding while ensuring that crucial geographical and topographical details are not lost in translation. In summary, the Pyramid Input mechanism within PCCAU-Net signifies a pioneering advancement in the field of high-resolution remote sensing image processing. Its inception marks a paradigmatic shift toward holistic multi-scale image interpretation, solidifying its potential as a prototype for future endeavors.

2.3. CoordConv Layer

CoordConv is an innovative convolutional variant that integrates spatial coordinates as additional channels, facilitating the direct interpretation of feature locations within a network [39]. PCCAU-Net judiciously adopts this technique as a cornerstone within its deep learning pipeline. Traditional convolutional operations prioritize translational invariance, serving as a rudimentary stage for feature abstraction [40]. However, this approach can sometimes act as a constraint when parsing high-resolution images replete with geographical information. Intricacies like topographic variances, road layouts, and other geographical features are deeply intertwined with their spatial coordinates. PCCAU-Net quickly surpassed this limitation by embedding spatial awareness into its architecture by implementing CoordConv. Importantly, this is not a mere capitulation to technological trends but an acknowledgment of its substantive contributions. In the initial stage of the model, CoordConv is invoked five times in sequence—first targeting the original size remote sensing imagery, and subsequently being deployed across four pyramid levels (P1, P2, P3, and P4). This design assures that, from the outset, location information at each spatial scale is deeply explored and integrated. This forward-thinking approach confers heightened acuity to PCCAU-Net during feature extraction, enabling a more intuitive grasp of spatial interrelations between entities beyond mere inference. This capability becomes particularly salient in decoding complex terrains, especially when certain spatial arrangements among features carry explicit geographical meanings. Overall, PCCAU-Net's strategic employment of CoordConv marks a paradigmatic shift in our handling of high-resolution remote sensing data, aiming for a more precise and intuitive extraction of spatial information. The schematic representation of the CoordConv structure is illustrated in Figure 2.

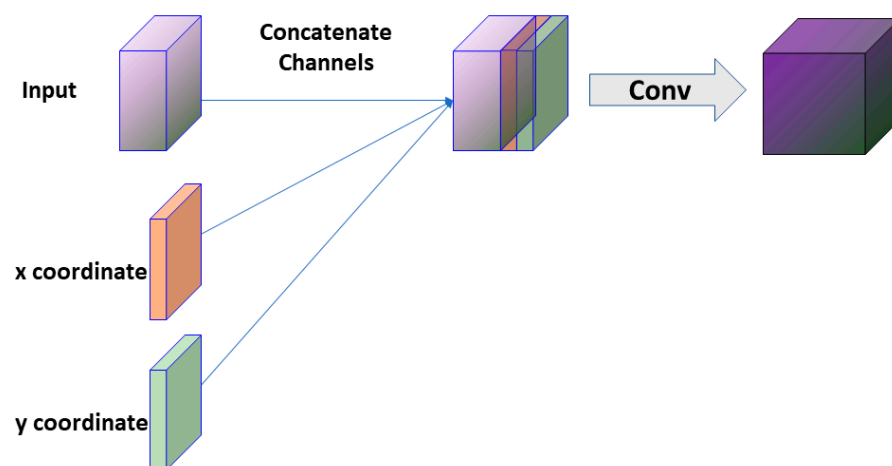


Figure 2. Schematic diagram of CoordConv operation details, where Input represents the input image, and x and y represent additional coordinate channels.

2.4. Dual-Input Cross-Attention (DCA) Module

In the intricate realm of remote sensing image analysis, developing ways of deftly and efficiently amalgamating features from disparate sources has become a cardinal research challenge. At the heart of PCCAU-Net is the Dual-Input Cross Attention (DCA) module, which answers this conundrum by adroitly harmonizing the strengths of deep learning and attention mechanisms. During the encoding phase, this module focuses on extracting and integrating features from two heterogenous pyramid pathways, ensuring a cohesive amalgamation of deep features. Even more critically, at the decoding phase, the DCA module serves as a skip connection, elegantly bridging deep and shallow features, and thereby achieving spatial and semantic coherence between them. The architecture of the DCA module is depicted in Figure 3.

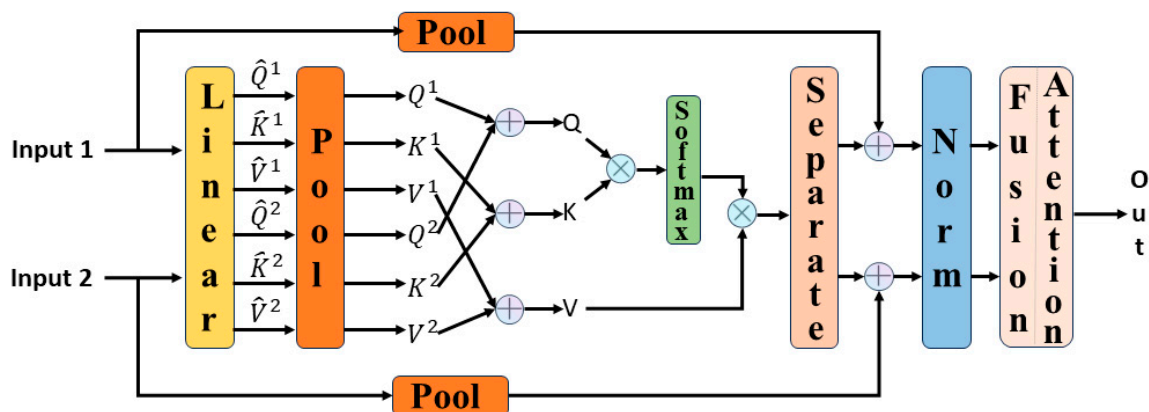


Figure 3. Schematic diagram of DCA module structure, where Input 1 and Input 2 represent two different inputs received by the module.

The intellectual underpinnings of the DCA module can be summarized as follows:

1. Feature encoding and interaction: the DCA module initially performs specialized convolution operations on both input features, generating query (Q), key (K), and value (V) representations. Ingeniously concatenated, these features complement and interact with each other in the subsequent attention stage.
2. Cross- Attention fusion: the module dynamically determines the importance of each feature by computing weights based on the juxtaposed Q and K. This process ensures that DCA can pinpoint and integrate salient information from varied origins.
3. Feature harmonization and dynamic fusion: the resultant features are first optimized via residual connections and subsequently subjected to layer normalization to ensure the continuity and stability of the information flow. Ultimately, the fusion attention mechanism calculates fusion weights and meticulously integrates Out_Input1 and Out_Input2, guaranteeing seamless spatial and semantic alignment.

In summary, the DCA module endows PCCAU-Net with an unprecedented capability for the effective interaction and integration of features across multiple scales and origins within a unified framework. This design ethos not only augments the model's feature representation capabilities but also brings about hitherto unmatched precision and efficiency in complex remote sensing image analysis tasks.

2.5. Fusion Attention Module

In the manifold applications of deep learning, particularly within the Dual-Input Cross Attention (DCA) module of the PCCAU-Net architecture, the nuanced and efficient fusion of multi-source information emerges as a critical determinant of model performance. To this end, we introduce a fusion attention mechanism within the DCA module, which not only orchestrates an organic integration of information but also assures the model's timely responsiveness to diverse information sources. This combination is then converted into a

two-dimensional space by a 1×1 convolutional layer to encode the importance of each source feature. Normalization via a Softmax operation ensures that the sum of feature weights at each position equals one. Armed with these dynamic weights, the module purposefully amalgamates the two input features, highlighting pivotal information in the final output. In summary, fusion attention serves as the DCA module's "arbiter", enabling the precise extraction and utilization of key information from multiple sources. This, in turn, elevates the sensitivity and accuracy in remote sensing image analysis, underscoring the considerable potential of deep learning in complex tasks. The schematic structure of the fusion attention module is illustrated in Figure 4.

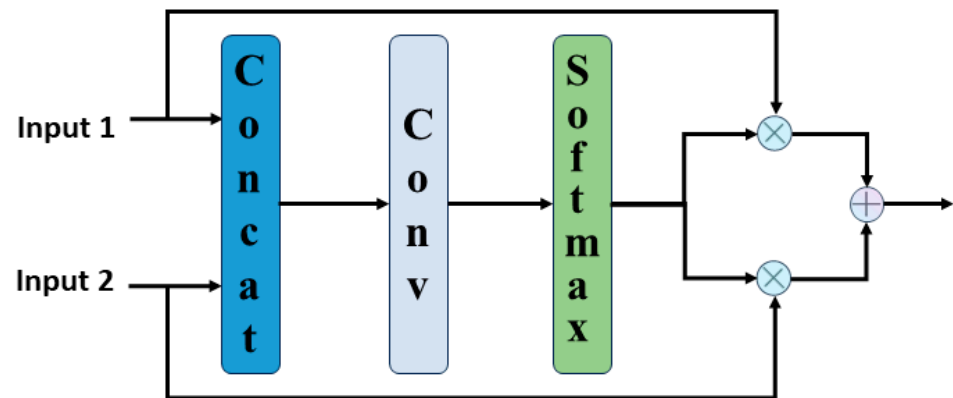


Figure 4. Schematic diagram of the fusion attention module structure.

2.6. Loss Function

The loss function is a function used to calculate the error between the predicted values of a deep learning model and its true values. A smaller loss indicates a smaller discrepancy between the model's output and the target values. During the model training process, the changes in the loss function provide a basis for the backpropagation process, guiding parameter optimization and helping the model's output to more closely approximate the target values. Dice loss is particularly suitable for addressing class imbalance issues. The Dice coefficient is used to measure the consistency between the predicted segmentation area and the actual segmentation area. The spatial relationship between positive samples and negative samples represented by Dice loss effectively handles the problem of sample imbalance [40]. However, when there is very little or no overlap between the predicted and true labels, the gradient of Dice loss can be difficult to handle. Binary Cross-Entropy (BCE) usually exhibits a stable gradient descent process, which aids in effective model training. Yet, when dealing with imbalanced data, BCE might cause the model to be biased towards the majority class and insufficient to capture the characteristics of the minority class. If only pixel-wise cross-entropy loss is used, the model may be biased towards predicting classes with more samples. The calculation formula for the Dice loss function and cross-entropy loss function is shown as follows:

$$\text{Dice Loss} = 1 - \frac{2|X \cap Y|}{|X| + |Y|} \quad (1)$$

$$\text{BCE Loss} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad (2)$$

In the formula 1, $|X|$ represents the the number of pixels in the predicted image, $|Y|$ represents the number of pixels in the true label, and $|X \cap Y|$ represents the intersection between the predicted image and the true label. In the formula 2, y represents the true pixel label value, y_i represents the predicted label pixel value by the model, and N represents the number of pixels.

The model in this study was a binary classification problem that performed pixel-wise classification to determine whether each pixel was a road or background. Roads constituted

approximately 10% of the total area in the remote sensing image. To balance the advantages and drawbacks of different loss functions, this paper combines Dice loss with Binary Cross-Entropy (BCE) to propose a new Combined Loss function. In the early stages of prediction, when the prediction results are significantly different from the true labels, the BCE loss can provide a considerable gradient to help the model converge quickly. As the prediction approaches the actual label, the gradient of the BCE loss decreases. In contrast, Dice loss can provide consistent gradients at this stage, leading to better optimization of the model. This feature helps to optimize the model globally and locally, improves the handling of class imbalance problems, and enhances the generalization ability and robustness of the model. The Combined Loss function is formulated as follows:

$$\text{Combined Loss} = 0.5 \times \text{BCE} + 0.5 \times \text{Dice} \quad (3)$$

3. Experiments and Results

In this section, we offer an in-depth analysis of the PCCAU-Net architecture, focusing particularly on its key constituents: the Pyramid Pathway Input, DCA (Dual-Input Cross Attention), and fusion attention modules. Initially, this section delineates the characteristics of the datasets employed and the experimental settings, thereby facilitating scientific reproducibility. Subsequently, we embark on a multifaceted quantitative and qualitative evaluation, pitting PCCAU-Net against various baseline models to obtain a comprehensive comparison. Furthermore, ablation studies elucidate the performance impact of each core component, substantiating both their efficacy and indispensability. The aim of this section is to furnish the reader with a holistic understanding of the capabilities and limitations of the proposed model.

3.1. Data Description

In the diverse landscape of deep learning and computer vision, the fidelity and precision of model performance are predominantly dictated by the quality of datasets and meticulously engineered preprocessing pipelines [41]. To rigorously assess the efficacy of our proposed model—especially its adaptability across multiple scales and contexts—we elected to utilize two industry-standard road extraction datasets: the Massachusetts Roads Dataset [42] and the DeepGlobe Road Dataset [43]. Beyond conventional preprocessing routines, we integrated a suite of data augmentation techniques—including image rotation, flipping, and color adjustments—to further enrich the training corpus and bolster model generalizability. Notably, the experimental design was tailored to the individual datasets, resizing images in the Massachusetts Roads Dataset to 512×512 pixels while retaining the native 1024×1024 -pixel dimensions in the DeepGlobe Road Dataset. This stratified approach was not solely predicated upon dataset specific characteristics, such as resolution and information density or computational efficiency and resource allocation; it also furnished a comprehensive evaluative framework for model performance at different scales. Training and testing on larger image dimensions afforded the enhanced capture of intricate road networks and topographical features, thereby augmenting the model's generalizability across multiple scales and scenarios. Through this holistic preprocessing and data augmentation strategy, our intent was to provide an exhaustive and insightful assessment of model performance under a plethora of experimental conditions, thereby lending robust support and precise limitations analysis for its real-world applications.

3.1.1. Massachusetts Roads Dataset

The Massachusetts Roads Dataset comprises high-resolution (1m per pixel) remote sensing imagery, with a total of 1171 images, each of 1500×1500 pixels, accompanied by meticulously annotated ground truth labels. The dataset is partitioned into 1108 training images, 14 validation images, and 49 test images, each subset replete with precise label information. To align with the experimental framework of this study, the original 1500×1500 -pixel images underwent preprocessing to yield 512×512 -pixel image blocks,

culminating in a corpus of 10,098 images. These images were further distributed into training, validation, and test sets at a ratio of 7:2:1.

3.1.2. DeepGlobe Road Dataset

Originating from the CVPR DeepGlobe 2018 Road Extraction Challenge, the DeepGlobe Road Dataset consists of 8570 high-resolution (0.5 m per pixel) images, each measuring 1024×1024 pixels. This collection is segmented into 6626 training images, 1243 validation images, and 1101 test images. In particular, only the binary labels for the training set were publicly accessible subsequent to the conclusion of the challenge; therefore, this study specifically opted for the 6626 training images and their corresponding labels for experimental implementation purposes. These images were also partitioned randomly in a 7:2:1 ratio.

3.2. Experimental Setup

To ensure the reliability and reproducibility of this study, this section offers a comprehensive description of crucial aspects of the experimental design. The experimental setup is bifurcated into two main components: the first part, “Computational Environment and Hyperparameters”, provides an exhaustive outline of the hardware specifications, operating system, development environment, and pivotal training hyperparameters. The second part, “Validation Metrics”, elaborates on the multiple criteria employed for assessing model performance. Together, these components form the scaffolding of the experimental design and confer robust assurance to the accuracy and consistency of subsequent results. By delineating these key parameters and evaluation criteria explicitly, we aim to furnish a thorough and meticulous experimental design that facilitates accurate replication by other researchers.

3.2.1. Training Environment and Hyperparameters

All experiments were conducted on a high-performance computer outfitted with an Intel® Core™ i7-11700 processor (with a base clock speed of 2.50 GHz), an Nvidia GeForce RTX 3060 graphics card, and 12GB of RAM. The operating system deployed was Windows 10, while the development environment was anchored on JetBrains PyCharm, version 2023. As the principal deep learning library, we opted for PyTorch (version 1.11.0), primarily due to its rich API set and exceptional computational performance. For optimization, the adaptive moment estimation (Adam) algorithm was utilized, the hyperparameters of which are delineated in Table 1. These parameter settings are carefully tuned to achieve a balance between optimal performance of model training and computational efficiency. Through this highly specialized experimental configuration and parameter adjustment, we strive to ensure the reliability and reproducibility of the research results.

Table 1. Hyperparameter settings in experimental training.

Hyperparameters	Value
Epoch	100
Batch Size	4 (Massachusetts Roads Dataset) 2 (DeepGlobe Road Dataset)
Initial Learning Rate	1×10^{-4}
ϵ	1×10^{-8}
Weight Decay	0
Beta 1	0.9
Beta 2	0.999

3.2.2. Verification Indicator

To systematically quantify and evaluate the performance of our road extraction model in classification tasks, we employed a confusion matrix as the principal evaluative metric.

This matrix delineates four key parameters: True Positives (TP), representing the accurate classification of pixels as road; True Negatives (TN), indicating correctly identified non-road pixels; False Positives (FP), specifying erroneously classified road pixels; and False Negatives (FN), marking road pixels incorrectly categorized as non-road. Guided by these fundamental metrics, we computed three principal evaluative indices:

Precision (P): this index gauges the model's accuracy in labeling pixels as roads (i.e., true positives). The formula employed is as follows:

$$P = \frac{TP}{TP + FP} \quad (4)$$

Recall (R): this is employed to assess the model's capacity to correctly identify pixels that are, in fact, roads (i.e., true positives). The computational formula is as follows:

$$R = \frac{TP}{TP + FN} \quad (5)$$

Intersection over union (IoU): this metric assesses the overlap between the model's predicted region and the actual ground truth, offering a robust measure of spatial accuracy. The computation is as follows:

$$IoU = \frac{TP}{TP + FP + FN} \quad (6)$$

Collectively, these indices provide a comprehensive and nuanced evaluation of model performance, focusing not only on identification accuracy but also on the model's ability to handle different classifications (i.e., true and false positives and negatives). Through this integrated evaluation framework, we aim to illuminate the model's strengths and weaknesses in road extraction tasks, thereby furnishing a holistic and in-depth understanding of its capabilities.

3.3. Comparative Experiment Results and Analysis

This chapter is devoted to an exhaustive inquiry into the practical utility of the PCCAU-Net model in the realm of road detection, achieved through systematic evaluation and analysis. To rigorously and comprehensively assess the performance of our model in the task of road extraction, this study engages a suite of benchmark models commonly employed in image segmentation, namely U-Net [16], Residual U-Net [18] (ResU-Net), DeepLab V3+ [22], and SDUNet [44]. These models are widely influential and recognized within this field as comparative benchmarks. U-Net, notable for its simplicity and efficiency, excels in small-scale datasets, but falls short in tackling complex, multi-scale challenges. In contrast, ResU-Net enhances convergence speed and accuracy by incorporating residual connections, although it risks overfitting in smaller datasets. DeepLab V3+ employs a sophisticated multi-module architecture, excelling in capturing multi-scale contextual information, albeit at the cost of elevated computational and storage overheads. SDUNet amplifies its capability to capture both local and global information through spatial attention mechanisms and dense connections; however, its efficacy may wane when dealing with highly imbalanced or heterogeneous data. These comparative models furnish a broad performance baseline and facilitate a nuanced understanding of the strengths and limitations of our proposed approach from multiple perspectives. These models are subjected to empirical validation using two datasets that present distinct characteristics and challenges: the Massachusetts Roads Dataset and the DeepGlobe Road Dataset. We aim to illuminate the exceptional efficacy and unique advantages of PCCAU-Net in the context of high-resolution remote sensing imagery for road detection.

3.3.1. Analysis of Massachusetts Roads Dataset Results

Within this section, we undertake a comprehensive evaluation of the PCCAU-Net model using the Massachusetts Roads Dataset. The assessment is bifurcated into qualitative and quantitative dimensions. Firstly, the qualitative analysis employs visual comparisons to elucidate how PCCAU-Net fares against other advanced models in road extraction tasks. Secondly, the quantitative facet deploys a set of standard metrics to numerically gauge the model's performance. The synthesis of these evaluations aims to delineate both the advantages and potential of PCCAU-Net in high-resolution remote sensing road detection tasks. Qualitative results are illustrated in Figure 5, while quantitative outcomes are tabulated in Table 2.

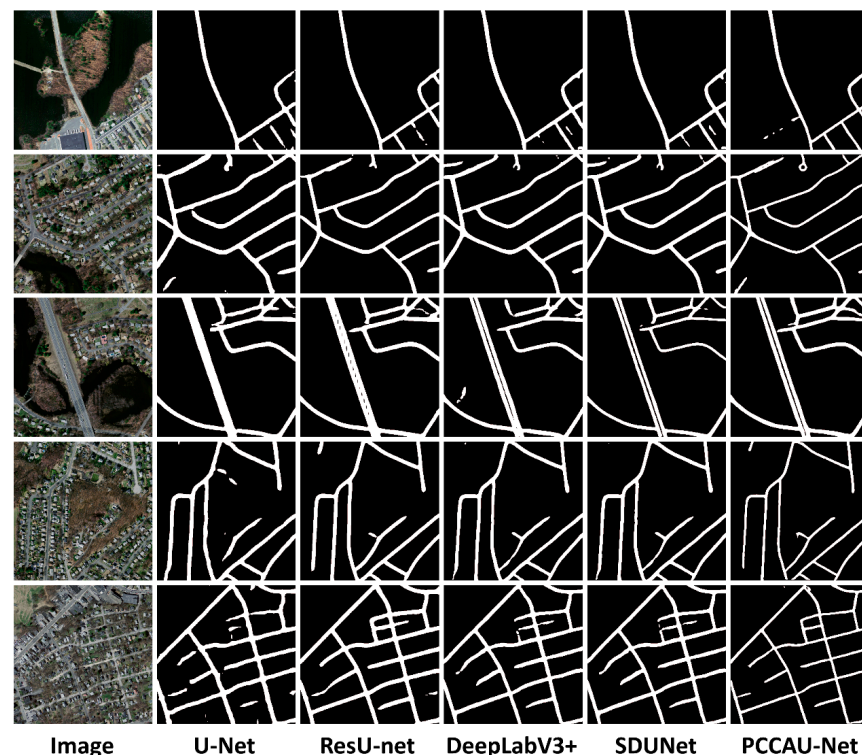


Figure 5. Extraction results of five models on the Massachusetts Roads Dataset.

Table 2. Quantitative Results Evaluation of Five Models on Massachusetts Roads Dataset.

Model	Precision (%)	Recall (%)	IoU (%)
U-Net	78.823	83.764	77.672
Resunet	79.297	83.253	77.902
DeeplabV3+	79.567	83.909	77.457
SDUNet	80.094	84.856	78.625
PCCAU-Net	83.395	84.076	78.783

Figure 5 showcases the performance of five models—U-Net, Residual U-Net (abbreviated as ResU-Net), DeepLab V3+, SDUNet, and our proposed PCCAU-Net—across five different scene captures from the Massachusetts Roads Dataset. These scenes represent a variety of intricate scenarios commonly encountered in remote sensing imagery. Scene 1 poses the particular challenge of severe spectral similarity between houses and roads, a hallmark issue in high-resolution remote sensing. While most models are susceptible to misclassifications in such contexts, PCCAU-Net demonstrates enhanced robustness. Scene 2 predominantly features a roundabout, which is traditionally a complicating factor in

remote sensing analyses. PCCAU-Net manages to differentiate it accurately, despite its complex morphology. Scene 3 illustrates the issue of delineating dual carriageways, a task at which conventional models often fail by mistaking them for a single road; PCCAU-Net identifies them distinctly. Scene 4 examines the model's capability to recognize detailed roads, and PCCAU-Net excels in capturing these intricacies. Finally, Scene 5 focuses on a major urban road with shadow occlusions, where PCCAU-Net exhibits remarkable resilience. Collectively, Figure 5, through its multi-angular and multi-layered exposition, authentically reflects the versatility and adaptability of PCCAU-Net in dealing with a multitude of complex remote sensing scenarios. This observation not only validates the significant advantages of the PCCAU-Net model in practical applications but also accentuates its broad applicability in high-resolution remote sensing road detection tasks.

Table 2 offers a comprehensive quantitative evaluation of the performance of five advanced architectures—namely, U-Net, ResU-Net, DeepLab V3+, SDUNet, and our proposed PCCAU-Net—on the Massachusetts Roads Dataset. This tabulation employs three pivotal metrics—precision, recall, and intersection-over-union (IoU)—to furnish a robust framework for assessing the relative efficacy of these models in the task of road extraction.

1. Precision: PCCAU-Net outperforms its peers with a precision rate of 83.395%, thereby indicating fewer false positives. The closest competitor, SDUNet, registers a precision of 80.094%, underscoring the substantial advantage of PCCAU-Net in minimizing identification errors.
2. Recall: While SDUNet achieves the highest recall rate at 84.856%, indicating its prowess in detecting the majority of true road segments; PCCAU-Net trails closely with a recall of 84.076%. This demonstrates a well-calibrated balance between precision and recall, offering a more nuanced view of road attributes.
3. IoU: PCCAU-Net leads with an IoU of 78.783%, narrowly eclipsing SDUNet's 78.625%. The superior IoU score for PCCAU-Net suggests its proficiency in delivering more accurate road segmentation, a critical aspect for applications demanding high precision.

In summary, PCCAU-Net manifests consistent superiority or near-top-tier performance across three key metrics, thereby substantiating its multifunctionality and robustness in high-resolution satellite imagery for road extraction. This exemplary performance validates the effectiveness of the attention mechanisms integrated within the DCA module, establishing PCCAU-Net as the preferred model for sophisticated road extraction tasks.

3.3.2. Analysis of DeepGlobe Roads Dataset Results

In this section, we delve into the experimental outcomes derived from deploying the PCCAU-Net model, along with several benchmark models, on the DeepGlobe Road Dataset. Similar to the approach adopted for the Massachusetts Roads Dataset, our evaluation incorporates both qualitative and quantitative analyses, illustrated in Figure 6 and Table 3, respectively.

Figure 6 presents road extraction results from five images, each embodying unique a challenge in diverse geographical settings: rural roads interspersed with natural scenery (Image 1), urban arterials suffering from significant occlusions (Image 2), suburban roads (Image 3), scenes with spectrally similar nonroad objects (Image 4), and highly intricate and dense urban road networks (Image 5).

Our analysis reveals that PCCAU-Net exhibits exceptional capabilities in road identification and extraction across these varying contexts, notably excelling in complex settings featuring occlusions and spectral similarities between roads and other elements. This superior performance can be attributed to the fusion attention mechanism embedded within the model's Dual-Input Cross Attention (DCA) module, which dynamically integrates information from diverse sources, thereby enhancing both the accuracy and robustness of the model for road extraction tasks across a spectrum of intricate and varied environments.

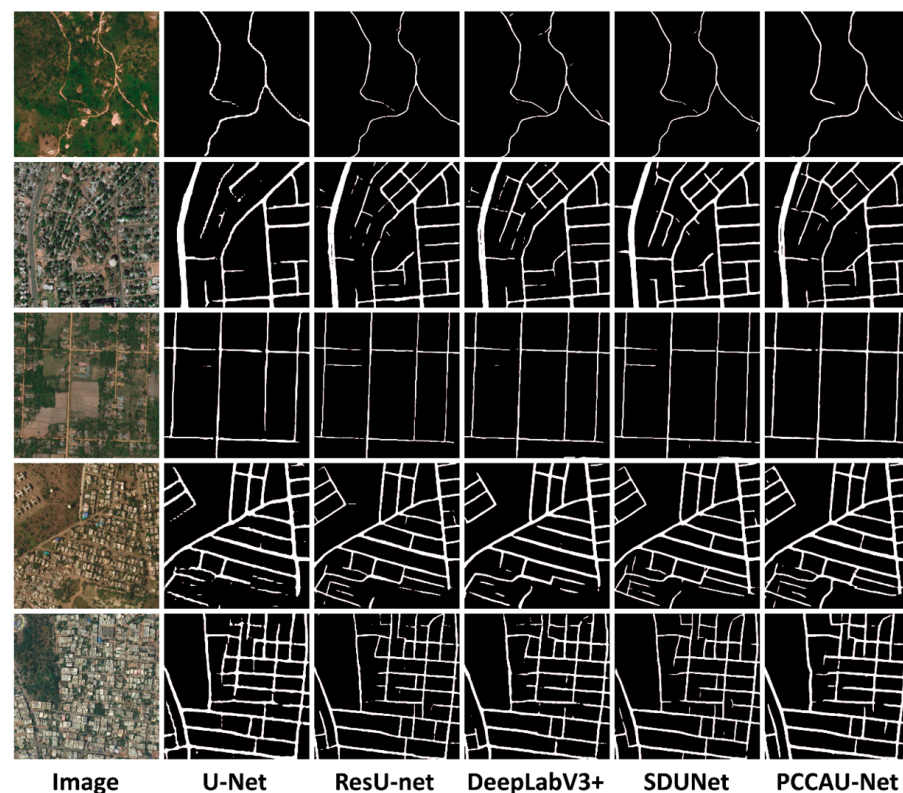


Figure 6. Extraction results of five models on the DeepGlobe Road Dataset.

Table 3. Quantitative Results Evaluation of Five Models on DeepGlobe Road Dataset.

Model	Precision (%)	Recall (%)	IoU (%)
U-Net	83.436	84.454	75.365
Resunet	86.098	86.085	77.819
DeeplabV3+	85.489	85.653	78.158
SDUNet	86.061	85.209	77.238
PCCAU-Net	87.424	88.293	80.113

The results presented in Table 3 unequivocally indicate that the PCCAU-Net model outperforms a range of benchmark models—including U-Net, ResU-Net, DeepLab V3+, and SDUNet—across key performance metrics on the DeepGlobe Road Dataset. Of particular salience is the marked enhancement PCCAU-Net achieves in terms of precision, recall, and Intersection-over-Union (IoU). Specifically, the model attains a precision of 87.424%, a recall of 88.293%, and an IoU as high as 80.113%, thereby eclipsing the performance of other models under comparable conditions. For instance, when juxtaposed with U-Net, PCCAU-Net surpasses the latter’s IoU by nearly five percentage points, thereby revealing its exceptional robustness and accuracy in confronting a multitude of geographical challenges, such as road occlusion and intricate road networks. Although the DeepLab V3+ model also performs commendably in the realms of precision and recall, its IoU score falls short of surpassing that of PCCAU-Net, further accentuating the latter’s superior performance in image segmentation tasks. This quantitative evidence underscores PCCAU-Net’s prowess in tackling complex road extraction tasks, largely attributable to its meticulously designed model, particularly the synergistic integration of Pyramid Pathway Input and DCA modules. Such innovative design elements not only bolster the model’s discriminative accuracy in intricate settings but also enhance its adaptability and resilience across varied geographical environments.

3.4. Ablation Experiments Results and Analysis

To rigorously evaluate both the efficacy of our proposed PCCAU-Net model and the contributions of its individual components, a series of ablation experiments were conducted. It is worth noting that while CoordConv convolution was employed in the first layer of the U-Net network, we forwent isolated ablation tests for this design, as its effectiveness had already been substantiated in the existing literature. The ablation tests were structured into five experimental setups as follows:

Setup a (U-Net): Serving as the baseline, this setup employs the canonical U-Net architecture. It provides a performance reference and forms the basis for comparison with other configurations.

Setup b (U-Net + Pyramid Pathway Input): Building upon the baseline U-Net, this setup incorporates Pyramid Pathway Input, designed to enhance the model's multi-scale feature processing for more precise road extraction.

Setup c (U-Net + DCA without fusion attention): In this configuration, a stripped-down DCA module, devoid of fusion attention, is integrated into the U-Net model. This is aimed at evaluating the fundamental contribution of the DCA module to the task of road extraction.

Setup d (U-Net + Complete DCA): This setup supplements the U-Net with a fully fledged DCA module, inclusive of fusion attention. The design is intended to appraise the comprehensive performance of the DCA module, and it serves as a comparative reference against Setup C to gauge the specific contributions of fusion attention.

Setup e (PCCAU-Net): This is the full configuration of our proposed model, which amalgamates both the Pyramid Pathway Input and the complete DCA module. The setup aims to demonstrate the optimal performance attainable when all design elements are integrated.

Through these five configurations, we not only methodically assess the effectiveness of both the Pyramid Pathway Input and the DCA module, but also gain comprehensive insights into their synergistic impact on the overall task of road extraction. The qualitative outcomes for these setups across two datasets are illustrated in Figure 7, and the quantitative findings are delineated in Table 4.

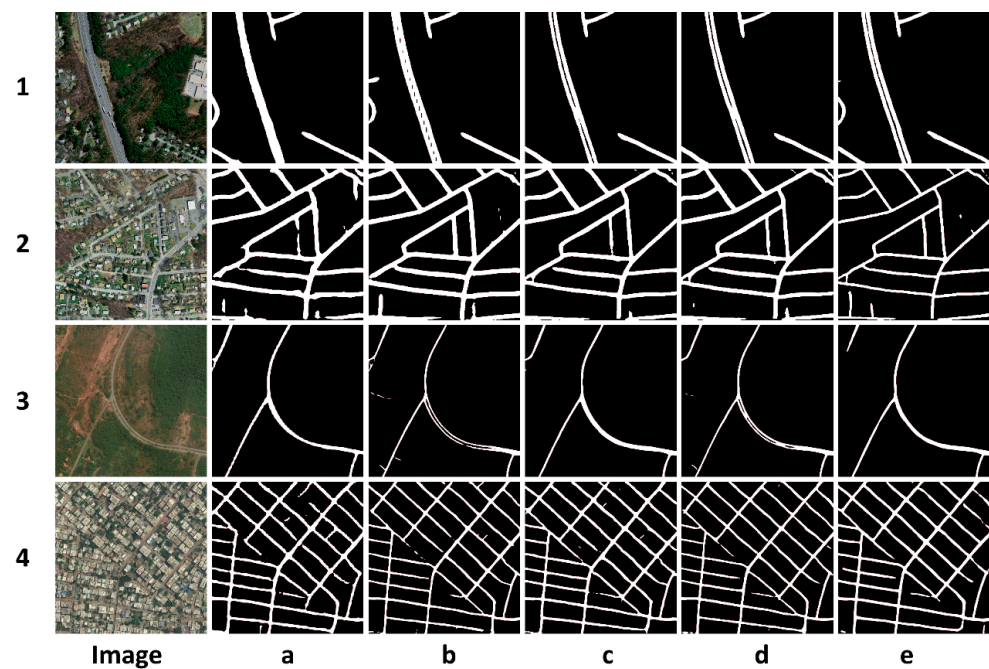


Figure 7. The test results of the ablation experiment scheme on two public road datasets, where images 1 and 2 are from the Massachusetts Roads Dataset, and images 3 and 4 are from the DeepGlobe Road Dataset.

Table 4. Quantitative results of five models of ablation experiments on two data.

Scheme	Massachusetts Roads Dataset			DeepGlobe Road Dataset		
	Precision (%)	ReCall (%)	IoU (%)	Precision (%)	ReCall (%)	IoU (%)
a	78.823	83.764	77.672	83.436	84.454	75.365
b	79.182	84.035	77.768	84.273	85.145	76.092
c	80.357	83.913	77.937	84.224	86.047	77.642
d	81.136	83.849	78.217	85.423	86.864	78.652
e	83.395	84.076	78.783	87.424	88.293	80.113

The letters “a, b, c, d, and e” in Figure 7 stand for experimental setups a, b, c, d, and e, respectively. In Figure 7, extraction outcomes from multiple approaches are depicted across four images, each presenting distinct challenges. The U-Net-based Scheme a offered a rudimentary performance across all images, particularly faltering in instances of intricate road interweaving and heterogeneity within the same spectral class. Scheme b, augmented with Pyramid Pathway Input, displayed nuanced improvements in detail extraction, but remained inadequate for handling same spectrum heterogeneity. Schemes c and d, both incorporating DCA modules, markedly ameliorated the extraction of complexly intertwined roads, with Scheme d showing a discernible edge in detail fidelity. Notably, Scheme e (PCCAU-Net) consistently outperformed others across all metrics, excelling particularly in scenarios involving complex road configurations and spectral variability.

In a quantitative analysis presented in Table 4, all five ablation study schemes were rigorously assessed on two distinct datasets—Massachusetts Roads Dataset and DeepGlobe Road Dataset. The letters a, b, c, d, and e shown in Table 4 have the same meanings as those in Figure 7. A detailed breakdown reveals that Scheme a (U-Net) registered the lowest aggregate performance across both datasets, especially in terms of Intersection-over-Union (IoU), displaying a conspicuous performance gap compared to other models. Scheme b (U-Net + Pyramid Pathway Input) achieved subtle yet consistent elevations across all evaluation metrics, underscoring the ancillary benefits of Pyramid Pathway Input. Scheme c (U-Net + DCA without functional attention) further optimized performance in metrics like IoU, suggesting that even without functional attention, DCA can bolster model efficacy. Scheme d (U-Net + full DCA) exhibited additional improvements across all metrics over Scheme c, highlighting the pivotal role of functional attention within the DCA module. Scheme e (PCCAU-Net) led the charts in all metrics across both datasets, showcasing the conspicuous advantages of the proposed Pyramid Pathway Input and a comprehensive DCA module.

Collectively, the incremental improvements from Scheme a to e not only manifest in a gradual performance escalation across various metrics but also unveil the critical contributions of Pyramid Pathway Input and DCA modules—particularly including functional attention—in optimizing model performance. These elements individually offer distinct advantages and, when synergistically integrated into the model, yield superlative performance, as conclusively validated in Scheme e. This comprehensive analysis robustly substantiates the theoretical and experimental underpinnings of this study.

3.5. Computational Efficiency

In the field of deep learning, the number of parameters and floating-point operations per second (FLOPs) serve as crucial measures for evaluating the complexity and computational demands of a model. A higher number of parameters might indicate a model’s ability to learn more intricate patterns, but it could also lead to overfitting and extended training time. A comparison of the computational efficiency of different methods is shown in Table 5. Our model, the PCCAU-Net, demonstrated balanced attributes with respect to both parameter quantity and FLOPs. Therefore, while significantly improving segmentation performance, the PCCAU-Net did not introduce an excessive number of parameters or

noticeably increase the training time. This demonstrates that our model not only prioritizes performance but also focuses on computational efficiency and model generalizability.

Table 5. Comparison of computational efficiency of different methods.

Model	Parameters (M)	FLOPs (GLOPS)
U-Net	29.97	5.63
Resunet	25.57	5.42
DeepLabV3+	5.88	6.60
SDUNet	25.90	62.30
PCCAU-Net	24.35	60.17

4. Discussion

In the current study, we successfully introduce PCCAU-Net, an advanced road extraction model, and subject it to comprehensive performance evaluation using two distinct road datasets from Massachusetts and DeepGlobe. When juxtaposed with leading algorithms like U-Net, ResU-Net, DeepLab V3+, and SDUNet, PCCAU-Net consistently outperforms them in key performance metrics such as precision, recall, and Intersection-over-Union (IoU). The efficacy of individual model components is investigated deeply through a series of ablation experiments. Notably, the Pyramid Pathway Input and DCA modules emerge as decisive factors in performance enhancement. Intriguingly, even a DCA module devoid of function attention contributes to model improvement, albeit less so than the fully equipped DCA, underscoring the critical role of function attention within the DCA framework. In real-world scenarios marked by spectral variations, shadow occlusions, and intricate road structures, PCCAU-Net demonstrates robust generalization capabilities. These claims are substantiated by qualitative results where the model reliably extracts roads, even in settings marred by complex details or significant occlusions.

In summary, this paper not only advances an efficient and robust algorithm for road extraction, but also corroborates its superiority and practicality through comparative and ablation studies. It thus offers a new, reliable toolset for automated road extraction and potentially for a broader range of remote sensing image analyses. However, the computational complexity of the model affects real-time applications, and in future research we will focus on optimizing the computational requirements while maintaining performance.

5. Conclusions

In this study, we present PCCAU-Net, an advanced deep learning architecture engineered for the efficient and accurate extraction of roads. Extensive evaluations on the Massachusetts Roads Dataset and the DeepGlobe Roads Dataset reveal that PCCAU-Net consistently outperforms extant cutting-edge methodologies, including U-Net, ResU-Net, DeepLab V3+, and SDUNet, across multiple evaluation metrics. The efficacy of PCCAU-Net is attributed to its innovative Pyramid Pathway Input and DCA modules, explicitly tailored to handle the complex and diverse scenarios inherent in remote sensing imagery. Ablation studies further validate the positive impact of these modules on the model's performance. However, it is imperative to acknowledge the model's computational complexity as a limitation, rendering it potentially unsuitable for real-time or high-throughput applications. Future work will be directed towards optimizing the model to meet these operational constraints. Overall, the findings not only advance the field of automated road extraction, but also offer valuable insights for additional applications in remote sensing image analysis.

Author Contributions: Conceptualization—X.X.; Methodology—C.R. and Y.Z.; Software—C.R., A.Y. and C.D.; Validation—A.Y. and J.L.; Formal analysis—X.X., Y.Z. and Y.L.; Resources—X.X. and C.R.; Data curation—X.X.; Writing—original draft, X.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research 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: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

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

References

- Oliveira, G.L.; Burgard, W.; Brox, T. Efficient deep models for monocular road segmentation. In Proceedings of the 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Daejeon, Republic of Korea, 9–14 October 2016; pp. 4885–4891. [\[CrossRef\]](#)
- Shamsolmoali, P.; Zareapoor, M.; Zhou, H.; Wang, R.; Yang, J. Road segmentation for remote sensing images using adversarial spatial pyramid networks. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 4673–4688. [\[CrossRef\]](#)
- Huan, H.; Sheng, Y.; Zhang, Y.; Liu, Y. Strip Attention Networks for Road Extraction. *Remote Sens.* **2022**, *14*, 4516. [\[CrossRef\]](#)
- Lyu, Y.; Bai, L.; Huang, X. Road segmentation using cnn and distributed lstm. In Proceedings of the 2019 IEEE International Symposium on Circuits and Systems (ISCAS), Sapporo, Japan, 26–29 May 2019; pp. 1–5.
- Lan, M.; Zhang, Y.; Zhang, L.; Du, B. Global context based automatic road segmentation via dilated convolutional neural network. *Inf. Sci.* **2020**, *535*, 156–171. [\[CrossRef\]](#)
- Ajayi, O.G.; Odumosu, J.O.; Samaila-Ija, H.A.; Zitta, N.; Adesina, E.A. *Dynamic Road Segmentation of Part of Bosso Local Government Area, Niger State*; Scientific & Academic: Rosemead, CA, USA, 2015.
- Sun, Z.; Zhou, W.; Ding, C.; Xia, M. Multi-resolution transformer network for building and road segmentation of remote sensing image. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 165. [\[CrossRef\]](#)
- Song, X.; Rui, T.; Zhang, S.; Fei, J.; Wang, X. A road segmentation method based on the deep auto-encoder with supervised learning. *Comput. Electr. Eng.* **2018**, *68*, 381–388. [\[CrossRef\]](#)
- Cheng, G.; Li, Z.; Yao, X.; Guo, L.; Wei, Z. Remote sensing image scene classification using bag of convolutional features. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1735–1739. [\[CrossRef\]](#)
- Blaschke, T.; Burnett, C.; Pekkarinen, A. Image segmentation methods for object-based analysis and classification. In *Remote Sensing Image Analysis: Including the Spatial Domain*; Springer: Berlin/Heidelberg, Germany, 2004; pp. 211–236.
- Lian, R.; Wang, W.; Mustafa, N.; Huang, L. Road extraction methods in high-resolution remote sensing images: A comprehensive review. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 5489–5507. [\[CrossRef\]](#)
- Furano, G.; Meoni, G.; Dunne, A.; Moloney, D.; Ferlet-Cavrois, V.; Tavoularis, A.; Byrne, J.; Buckley, L.; Psarakis, M.; Voss, K.-O. Towards the use of artificial intelligence on the edge in space systems: Challenges and opportunities. *IEEE Aerosp. Electron. Syst. Mag.* **2020**, *35*, 44–56. [\[CrossRef\]](#)
- Paisitkriangkrai, S.; Sherrah, J.; Janney, P.; Van Den Hengel, A. Semantic labeling of aerial and satellite imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2016**, *9*, 2868–2881. [\[CrossRef\]](#)
- 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.
- Sherrah, J. Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery. *arXiv* **2016**, arXiv:1606.02585.
- Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III* 18; Springer International Publishing: New York, NY, USA, 2015. [\[CrossRef\]](#)
- 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 (CVPR), Las Vegas, NV, USA, 27–30 June 2016. [\[CrossRef\]](#)
- Diakogiannis, F.I.; Waldner, F.; Caccetta, P.; Wu, C. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS J. Photogramm. Remote Sens.* **2020**, *162*, 94–114. [\[CrossRef\]](#)
- Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv* **2014**, arXiv:1412.7062.
- 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. [\[CrossRef\]](#) [\[PubMed\]](#)
- Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
- Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.

23. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep High-Resolution Representation Learning for Human Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5693–5703.
24. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. [\[CrossRef\]](#)
25. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceeding of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
26. 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. [\[CrossRef\]](#)
27. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
28. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. Ccnet: Criss-cross attention for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 603–612.
29. Cao, X.; Zhang, K.; Jiao, L. CSANet: Cross-Scale Axial Attention Network for Road Segmentation. *Remote Sens.* **2023**, *15*, 3. [\[CrossRef\]](#)
30. Zhang, Z.; Miao, C.; Liu, C.; Tian, Q.; Zhou, Y. HA-RoadFormer: Hybrid Attention Transformer with Multi-Branch for Large-Scale High-Resolution Dense Road Segmentation. *Mathematics* **2022**, *10*, 1915. [\[CrossRef\]](#)
31. Tong, Z.; Li, Y.; Zhang, J.; He, L.; Gong, Y. MSFANet: Multiscale Fusion Attention Network for Road Segmentation of Multispectral Remote Sensing Data. *Remote Sens.* **2023**, *15*, 1978. [\[CrossRef\]](#)
32. Alshaikhli, T.; Liu, W.; Maruyama, Y. Simultaneous Extraction of Road and Centerline from Aerial Images Using a Deep Convolutional Neural Network. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 147. [\[CrossRef\]](#)
33. Zhu, Q.; Zhang, Y.; Wang, L.; Zhong, Y.; Guan, Q.; Lu, X.; Zhang, L.; Li, D. A Global Context-aware and Batchin dependent Network for road extraction from VHR satellite imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *175*, 353–365. [\[CrossRef\]](#)
34. Shao, Z.; Zhou, Z.; Huang, X.; Zhang, Y. MRENet: Simultaneous Extraction of Road Surface and Road Centerline in Complex Urban Scenes from Very High Resolution Images. *Remote Sens.* **2021**, *13*, 239. [\[CrossRef\]](#)
35. Rong, Y.; Zhuang, Z.; He, Z.; Wang, X. A Maritime Traffic Network Mining Method Based on Massive Trajectory Data. *Electronics* **2022**, *11*, 987. [\[CrossRef\]](#)
36. Li, Z.; Fang, C.; Xiao, R.; Wang, W.; Yan, Y. SI-Net: Multi-Scale Context-Aware Convolutional Block for Speaker Verification. In Proceedings of the 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Cartagena, Colombia, 13–17 December 2021; pp. 220–227. [\[CrossRef\]](#)
37. Kang, L.; Riba, P.; Rusiñol, M.; Fornés, A.; Villegas, M. Pay attention to what you read: Non-recurrent handwritten text-line recognition. *Pattern Recognit.* **2022**, *129*, 108766. [\[CrossRef\]](#)
38. Qin, N.; Tan, W.; Ma, L.; Zhang, D.; Guan, H.; Li, J. Deep learning for filtering the ground from ALS point clouds: A dataset, evaluations and issues. *ISPRS J. Photogramm. Remote Sens.* **2023**, *202*, 246–261. [\[CrossRef\]](#)
39. Liu, R.; Lehman, J.; Molino, P.; Petroski Such, F.; Frank, E.; Sergeev, A.; Yosinski, J. An intriguing failing of convolutional neural networks and the coordconv solution. *Adv. Neural Inf. Process. Syst.* **2018**, *31*. [\[CrossRef\]](#)
40. Mena, J.B. State of the art on automatic road extraction for GIS update: A novel classification. *Pattern Recognit. Lett.* **2003**, *24*, 3037–3058. [\[CrossRef\]](#)
41. Emam, Z.; Kondrich, A.; Harrison, S.; Lau, F.; Wang, Y.; Kim, A.; Branson, E. On the state of data in computer vision: Human annotations remain indispensable for developing deep learning models. *arXiv* **2021**, arXiv:2108.00114.
42. Mnih, V. *Machine Learning for Aerial Image Labeling*; University of Toronto: Toronto, ON, Canada, 2013.
43. 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, 18–23 June 2018; pp. 172–181.
44. SDUNet: Road Extraction via Spatial Enhanced and Densely Connected UNet. Available online: <https://www.sciencedirect.com/science/article/pii/S0031320322000309> (accessed on 9 March 2023).

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.