International Journal of Geo-Information

MDPI

Article

An Efficient Algorithm for Extracting Railway Tracks Based on Spatial-Channel Graph Convolutional Network and Deep Neural Residual Network

Yanbin Weng [1,*], Meng Xu [1], Xiahu Chen [1,2], Cheng Peng [1], Hui Xiang [1], Peixin Xie [1] and Hua Yin [1]

[1] College of Computer Science, Hunan University of Technology, Tianyuan District, Zhuzhou 412007, China; m22077500024@stu.hut.edu.cn (M.X.); chenxh@hut.edu.cn (X.C.); pengcheng@hut.edu.cn (C.P.); m23085400021@stu.hut.edu.cn (H.X.); m23077500010@stu.hut.edu.cn (P.X.); 21405700636@stu.hut.edu.cn (H.Y.)

[2] Zhuzhou Taichang Electronic Information Technology Co., Ltd., Zhuzhou 412007, China

* Correspondence: wengyb@hut.edu.cn

Abstract: The accurate detection of railway tracks is essential for ensuring the safe operation of railways. This study introduces an innovative algorithm that utilizes a graph convolutional network (GCN) and deep neural residual network to enhance feature extraction from high-resolution aerial imagery. The traditional encoder–decoder architecture is expanded with GCN, which improves neighborhood definitions and enables long-range information exchange in a single layer. As a result, complex track features and contextual information are captured more effectively. The deep neural residual network, which incorporates depthwise separable convolution and an inverted bottleneck design, improves the representation of long-distance positional information and addresses occlusion caused by train carriages. The scSE attention mechanism reduces noise and optimizes feature representation. The algorithm was trained and tested on custom and Massachusetts datasets, demonstrating an 89.79% recall rate. This is a 3.17% improvement over the original U-Net model, indicating excellent performance in railway track segmentation. These findings suggest that the proposed algorithm not only excels in railway track segmentation but also offers significant competitive advantages in performance.

Keywords: graph convolution; railroad track extraction; residuals; attention mechanism

Citation: Weng, Y.; Xu, M.; Chen, X.; Peng, C.; Xiang, H.; Xie, P.; Yin, H. An Efficient Algorithm for Extracting Railway Tracks Based on Spatial-Channel Graph Convolutional Network and Deep Neural Residual Network. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 309. https://doi.org/10.3390/ijgi13090309

Academic Editor: Wolfgang Kainz

Received: 18 July 2024
Revised: 18 August 2024
Accepted: 27 August 2024
Published: 29 August 2024

1. Introduction

China's transportation sector development has consistently ranked among the top globally [1]. The segmentation and extraction technology for railway tracks can optimize the utilization of railway lines, reduce resource wastage, and improve energy efficiency. Additionally, it can provide valuable information for land use planning, emergency response, and monitoring the evolution of railway networks [2]. Due to the movement and occlusion of train carriages on the tracks, traditional detection methods struggle to accurately identify and track railway tracks. This not only affects the safety of railway operations but also impacts the efficiency of railway line usage and the rational allocation of resources. Scientifically and reasonably analyzing the distribution of railway lines can provide strong support for formulating regional economic policies, promoting coordinated development among different areas, and fostering comprehensive, balanced, and sustainable economic growth in China [3].

In recent years, deep learning technology has gained recognition for its outstanding performance and generalization capabilities, and its application in road extraction has become increasingly widespread [4]. Railway track information shares similarities with road information, allowing for extraction using similar principles. Semantic segmentation of railway tracks is a highly challenging task, where each pixel belonging to the track

must be labeled as railway, while the remaining pixels are labeled as background, thus constituting a binary semantic segmentation problem [5]. Compared to general semantic segmentation tasks, railway track segmentation has unique challenges and complexities.

Current mainstream road extraction methods mainly classify texture, geometric, and photometric features [6], but these methods are primarily tailored for roads, presenting numerous challenges for railway track semantic segmentation. In the field of image processing, including image classification, scene recognition, object detection, and semantic segmentation, advanced convolutional neural networks (CNNs) have achieved remarkable success [7]. At the end of 2012, the AlexNet18 model was proposed as one of the notable methods for promoting CNNs in computer vision [8], followed by the emergence of many representative deep networks. Fully convolutional networks (FCNs) achieve pixel-level prediction through techniques such as convolution, upsampling, and skip connections, being the first to offer a fully end-to-end supervised and pre-trained technology [9]. However, FCNs suffer from a loss of spatial resolution due to pooling layers, which can lead to inaccurate edge detection, a critical issue in railway track segmentation.

Advanced models like U-Net, SegNet, and DeepLab have shown satisfactory results in road extraction but often misclassify non-road areas as complex regions, resulting in many false negatives. Additionally, due to the pooling layers in FCNs, while increasing the receptive field, the models lose edge positional information. To fully utilize the feature information in images, the U-Net series employs skip connection structures [10], achieving multi-scale image information fusion and improving performance. VGGNet [11] proposed the idea of constructing deep models by repeatedly applying simple basic blocks, demonstrating that the depth of the network is a key component of the exceptional performance of deep learning models. However, VGGNet's architecture can be computationally expensive and may not be optimal for high-resolution semantic segmentation where computational efficiency is crucial. Wang et al. [12] used a finite state machine (FSM) and deep neural networks (DNNs) to extract road classes from high-resolution aerial and Google Earth images, extracting roads through training and tracking steps, though this method could not address occlusion issues. Zhou et al. adopted an encoder–decoder structure and proposed D-LinkNet [13], which enhanced the overall receptive field of the model using dilated convolutions without reducing the resolution of the feature maps. While D-LinkNet improves the receptive field, it can still struggle with small object detection and precise boundary localization.

Graph convolutional networks (GCNs) expand convolution operations to a broader graph domain, inheriting pooling and convolution operations from CNNs. This allows for the combination of local feature information, the removal of redundant information, and the enhancement of edge and small object extraction accuracy. Kipf and Welling proposed the semi-supervised GCN model, which has become a foundational method in the application of GCNs, showing that GCNs can effectively capture spatial relationships within data [14]. The proposed method builds on this concept, using GCNs not only for capturing spatial relationships but also for enhancing the model's capability to distinguish between railway tracks and visually similar features, such as train carriages. Lu et al. [15] were the first to use graph convolution in combination with fully convolutional neural networks to address the semantic segmentation challenge, resulting in a 1.34% performance improvement. Nevertheless, this approach focuses mainly on general segmentation tasks and does not specifically address the extraction of elongated shapes, such as railroad tracks or roads.

Chen et al. [16] proposed a GCN algorithm integrating distance and direction by constructing a dynamic neighborhood graph through the similarity matrix of point clouds, thereby better capturing the local geometric features of point clouds. Although this approach performs well in point cloud segmentation, it is not suitable for 2D image segmentation tasks with pixel-level classification. In our model, the use of GCN focuses on enhancing spatial relationships in high-resolution images, which is more conducive to distinguishing railroad tracks from train cars. Yu et al. [17] introduced a graph convolution method based

on GraphSAGE, utilizing directed graphs and dual graphs to construct the graph structure, defining graph descriptions from different perspectives to effectively identify spatial cognitive features. This approach primarily focuses on high-level spatial features and cannot effectively capture fine-grained details. Despite the extensive application and superior performance of graph convolution structures in the field of image recognition, their use in pixel classification and road extraction is rare. MobileNetv2 [18] introduced a residual structure with depthwise separable convolutions and linear bottlenecks, significantly reducing the number of parameters and enhancing the network's representational capacity. Dai et al. [19] proposed a network structure composed of concatenated pointwise convolutions and multi-scale depthwise convolutions, which extract multi-scale spatial features while maintaining a lightweight architecture. Despite its ability to handle multi-scale features, this method may struggle with occlusions and complex background environments, which are common in railway track images. Huang et al. [20] proposed an end-to-end depthwise separable U-shaped convolutional network, which utilizes large kernel convolutions to obtain a larger receptive field for feature extraction, thereby effectively improving the performance of multi-scale medical segmentation. Graph convolutional LSTM [21] is well suited for scenarios where temporal dynamics are critical, but in my static imagery context, the temporal aspect is not a factor. CycleGAN [22] is designed for unpaired image-to-image translation, which does not directly address the precise semantic segmentation requirements for railway tracks. Therefore, the selected GCN-based approach, which excels in capturing spatial relationships and enhancing boundary delineation, is more appropriate for specific applications.

In high-resolution images, railway tracks, although occupying a small proportion of the entire image, typically span the entire image, and their texture features can easily be confused with the surrounding background environment. Unlike road information, railway tracks are often covered by train carriages, making the correct distinction between ordinary tracks and train carriages a critical issue in track extraction methods. Additionally, multiple tracks intersecting to form switches increase the complexity of the topological structure. These factors make the extraction of railway tracks from aerial images challenging and weaken the applicability of many semantic segmentation methods. Given the aforementioned issues and challenges, to improve the segmentation performance of railway tracks, enhance the model's feature extraction capability, and preserve image detail features, we propose a railway track extraction model that integrates a graph convolutional network and deep residual networks. The main contributions of this paper are as follows:

1. Establishment of a segmentation dataset: High-resolution images of train stations in southern China were obtained via drone photography, annotated, calibrated, and manually corrected using Labelme 3.16.7. Data augmentation techniques such as cutting, rotating, and flipping were employed to expand the dataset, resulting in a railway dataset comprising 13,285 images.
2. Introduction of deep neural residual network and graph convolutional network: The deep neural residual network consists of multiple residual structures in series and parallel, combined with depthwise separable convolutions and inverted residual structures, enabling the sharing of multi-scale convolution kernel parameter information and addressing the issue of train carriage occlusion. The graph convolution structure separately processes spatial and channel features, enhancing the representation of track features.
3. Optimization of the overall structure: The original RELU activation function was replaced with PRELU, and the number of model layers was reduced to ensure segmentation efficiency. The scSE attention mechanism module was added to the encoder and decoder to suppress unimportant features, reduce noise interference, and enhance foreground response, thereby improving segmentation accuracy and robustness.

2. Materials and Methods

2.1. Algorithm Flow

To maximize the useful information in railway images and improve the accuracy of segmenting the target area, this study consists of five main steps: dataset initialization, data preprocessing, training the railway track segmentation network, postprocessing, and model evaluation. The process is visualized in Figure 1.

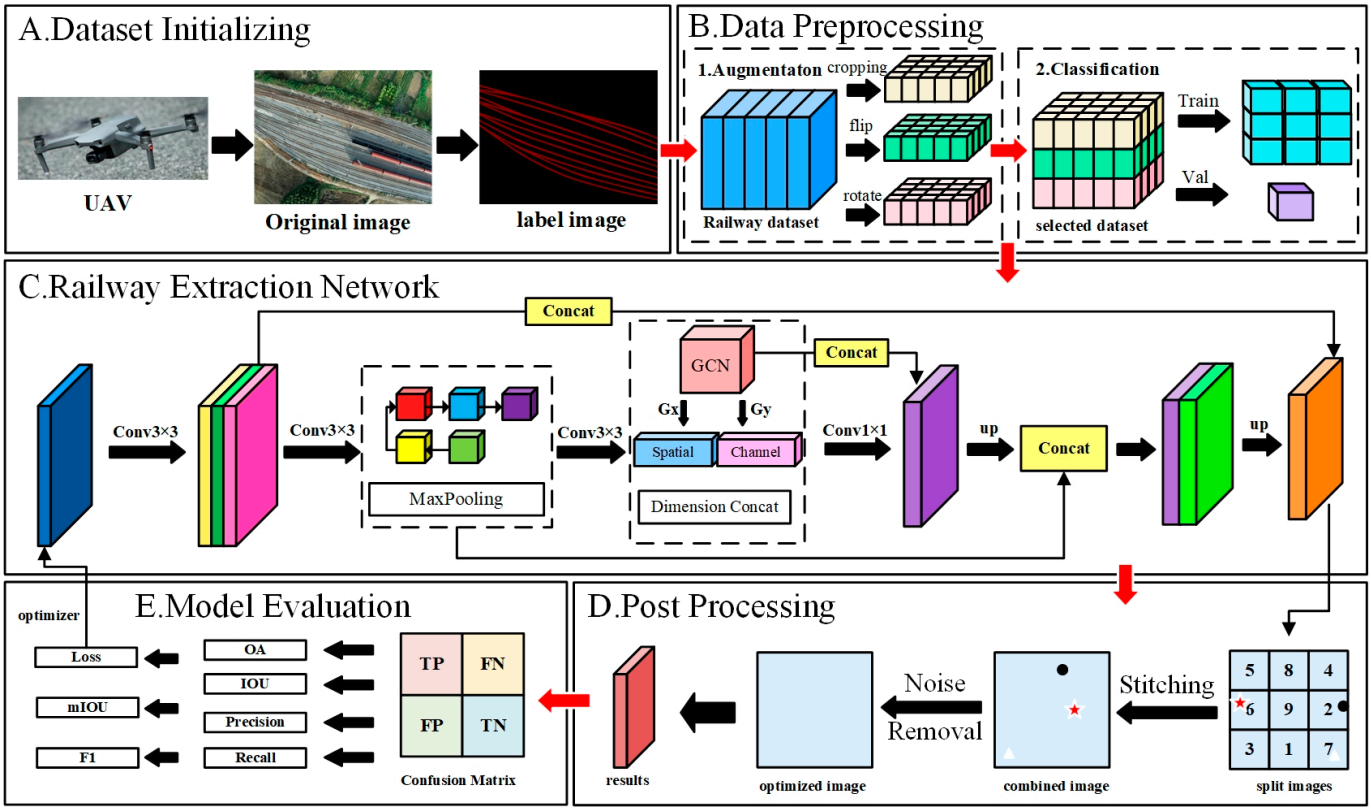


Figure 1. Flowchart of the algorithm proposed in this article.

Given the limited quantity and poor quality of existing railway datasets, this study created its aerial image dataset. The dataset was formed by capturing images of railway areas using drone aerial photography and then labeling the images to mark railway tracks in red. The data were divided using a sliding window algorithm [23] and underwent operations such as rotation and translation to further enhance the dataset. The dataset was then split into training and validation sets, which were used for training the railway segmentation network. Subsequently, the trained weights were used to evaluate pre-extracted images, producing black-and-white mask images depicting the extracted railway tracks in white.

During the railway track segmentation process, challenges often arose regarding the low accuracy of the extraction results. Given the high resolution and complex information of the images used in this study, obtaining satisfactory extraction results was particularly challenging. To address this, images were segmented into smaller blocks and augmented to increase the dataset size, with the extracted results reassembled in the process. Subsequently, morphological algorithms [24] were employed to eliminate noise and refine the extraction results. Finally, model performance was evaluated using a confusion matrix and standard metrics such as MIoU, accuracy, and F1 score. Loss values were iteratively fed back into the network structure during training to obtain optimal weight values.

2.2. Improved U-Net Network Structure

The U-Net architecture is widely recognized in the field of image segmentation for its exceptional performance [25]. This study utilizes the conventional U-Net framework, which employs an encoder–decoder structure with skip connections to create a comprehensive image segmentation model. The model consists of a three-layer encoder and corresponding decoders, forming a seamless network flow from input to output. This three-layer depth was selected to strike a balance between computational efficiency and effective feature extraction. This ensures that important image features are captured while still maintaining a manageable level of model complexity. The encoder section conducts feature extraction using convolutional layers with a 3×3 kernel size and a stride of 2 for efficient downsampling. It also includes 2×2 max pooling operations and standard bottleneck residual blocks to enhance the model's ability to capture image features. Batch normalization (BN) and parametric rectified linear unit (PReLU) activation functions are applied after each convolutional operation to enhance the model's nonlinear expression capability and training stability. The input layer of the encoder is designed with 3 channels to accommodate color image inputs. To mitigate potential feature loss during consecutive downsampling, this study cleverly integrates multiple residual blocks at the end of the encoder. These blocks ensure that high-level spatial features in the image are retained even after downsampling by introducing residual connectivity. This connectivity helps the model maintain accurate boundary localization by still being able to retain accurate spatial information at a deep level, preventing the loss of boundary information.

In both the residual chain and decoders, this study introduces the spatial channel squeeze-and-excitation module (scSE), which integrates channel and spatial dimension information to enrich the representation of deep semantic features. The scSE mechanism enables the model to focus on important spatial features while suppressing irrelevant background noise. Furthermore, this research incorporates a graph convolutional network (GCN) at the lower layers of the model to capture deep image features and their interrelationships. GCN enhances the model's ability to understand and process spatial relationships in an image by constructing a graph structure that captures the spatial relationships between different regions in an image. This structure helps to accurately delineate different regions in an image, especially boundaries in complex structures, which are essential for accurate boundary detection. By constructing a graph structure, the model can gather global contextual information and use it as node features, effectively enhancing its ability to recognize and segment complex image structures. The network architecture diagram is shown in Figure 2.

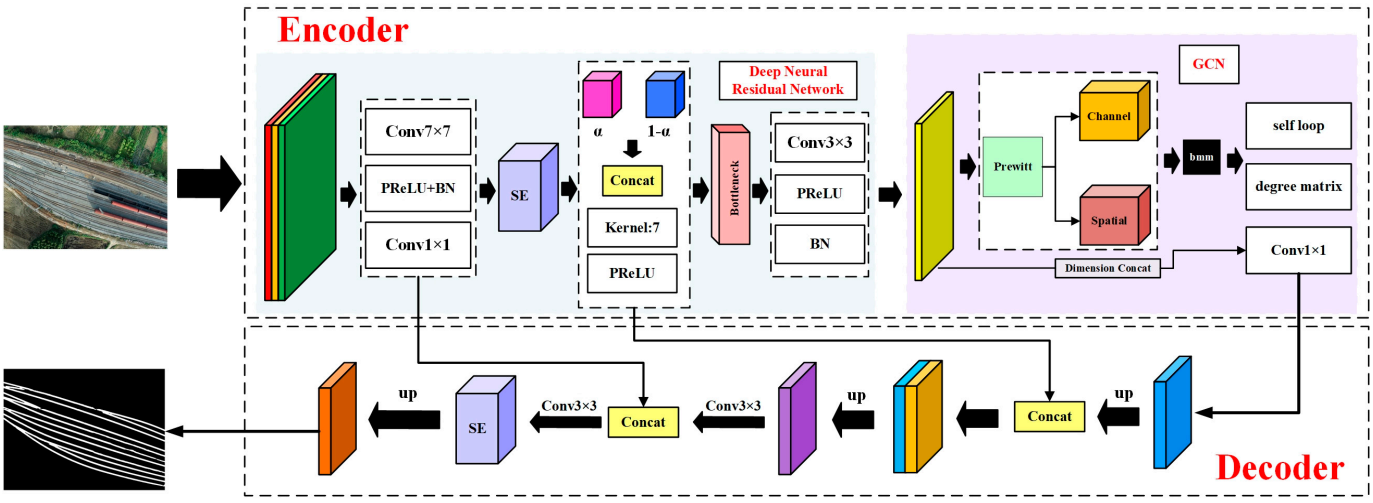


Figure 2. Improved U−Net network structure diagram.

In the decoder component of the model, three transposed convolutional layers are utilized to progressively upscale the feature maps with a stride of 2. After each upscaling operation, two consecutive 3×3 convolutional layers are employed to further enhance the feature representation. The resultant outputs of these convolutional layers are subject to normalization via batch normalization to promote stabilization of the learning process and expedite convergence. Following this, a parametric rectified linear unit (PReLU) activation function is invoked to introduce nonlinearity and augment the model's expressive capacity. To bolster the model's generalization and performance, a spatial channel attention module is integrated immediately preceding the final upscaling layer, thereby effectively refining and fortifying the feature representation to optimize the conclusive prediction outcomes. Furthermore, a skip connection mechanism is deployed to amalgamate feature maps from corresponding depths in the encoder with higher-level features in the decoder. This fusion approach serves to not only conserve rich semantic information but also guarantee the effective propagation of detailed information, thereby establishing an equilibrium between semantics and details in the decoder's output. Consequently, this endeavor culminates in an augmentation of the model's predictive precision and efficacy.

2.3. Deep Neural Residual Network

The core objective of feature extraction is to identify effective methodologies for extracting comprehensive and structured information. Conventional semantic segmentation networks depend exclusively on ordinary convolutions for feature extraction. However, as network depth increases, challenges such as convergence issues and gradient vanishing during training are predisposed to arise. Architectures featuring residual connections can circumvent gradient vanishing issues [26] and attain superior accuracy.

A solitary-instance local learning approach proves insufficient for acquiring the most precise and useful feature information. Consequently, this study endeavors to augment the receptive field of feature propagation by amalgamating multiple residual blocks to facilitate information exchange across larger spatial extents. While this strategy widens the perception field by stacking modules, it may exhibit heightened sensitivity to comparable objects in varied positions. Nevertheless, the eminence of residual structures ensures the non-occurrence of gradient convergence issues, even with a notably amplified module depth.

Residual connections characterize outputs primarily as a linear combination of inputs and nonlinear transformations. The incorporation of three types of residual blocks in both series and parallel refines the training format of high-dimensional features, thereby enriching the precision of corresponding information extraction. The enhanced residual structure is delineated in Figure 3.

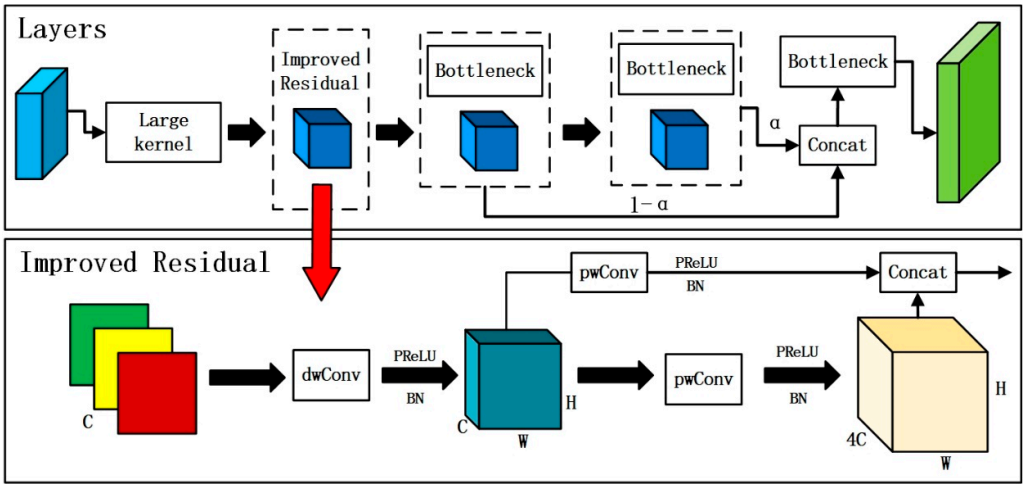


Figure 3. Deep neural residual network diagram.

The bottleneck is a standard residual bottleneck that utilizes 3 × 3 convolutions combined with 1 × 1 convolutions. The primary concept of the large kernel block involves integrating a 7 × 7 large kernel convolution with residual connections, incorporating an inverted bottleneck design for convolutions, and replacing the traditional ReLU function with a Gaussian error linear unit (GELU). GELU combines nonlinearity and stochastic regularization by weighting inputs based on their magnitude, outperforming ReLU and ELU in various tasks. The core component of the improved residual block is depthwise separable convolution, which was popularized by MobileNetv2 and MobileViT [27]. Depthwise separable convolution [28] combines depthwise and pointwise convolutions instead of a full standard convolution operation. Depthwise convolution extracts spatial dimension information, followed by pointwise convolution that mixes spatial and channel separations. Compared to standard convolutions, depthwise convolutions effectively reduce network parameters and computational load.

To capture more feature information, the depth of the entire module is increased. In the improved residual block, depth convolutions of large kernel sizes are used to extract global information for each channel, followed by residual connections. To fully integrate spatial and channel information, two pointwise convolutions are applied after depthwise convolutions, designed with inverted bottleneck configurations, setting the hidden dimensions between the two pointwise convolution layers to four times the input width. The inverted bottleneck design has been further extended in ConvNeXt. The extended hidden dimensions comprehensively blend global spatial dimension information extracted by depthwise convolutions. The model employs a total of 20 residual blocks, strategically distributed throughout its architecture. It starts with three large kernel blocks, followed by one improved residual block. In the subsequent layers, four bottleneck blocks are combined with one improved residual block, followed by six bottleneck blocks paired with another improved residual block. The outputs from these layers are weighted and concatenated before being passed through three additional bottleneck blocks, resulting in the final output. Additionally, PReLU activation and BN layers are used after each convolution. The formula for the improved residual block is defined as follows:

$$\mathrm{Imp}_{\mathrm{Res}}(i) = \mathrm{BN}\big(\mathrm{PRELU}\{\mathrm{dw}\big(\mathrm{Imp}_{\mathrm{Res}}(i-1)\big)\}\big) + \mathrm{Imp}_{\mathrm{Res}}(i-1) \quad (1)$$

$$\mathrm{Imp}_{\mathrm{Res}}(i+1) = \mathrm{BN}\big(\mathrm{PRELU}\{\mathrm{pw}\big(\mathrm{Imp}_{\mathrm{Res}}(i)\big)\}\big) \quad (2)$$

$$\mathrm{Imp}_{\mathrm{Res}}(i+2) = \mathrm{BN}\big(\mathrm{PRELU}\{\mathrm{pw}\big(\mathrm{Imp}_{\mathrm{Res}}(i+1)\big)\}\big) \quad (3)$$

where $\mathrm{Imp}_{\mathrm{Res}}$ denotes the output of the improved residual block, dw denotes deep convolution, and pw denotes pointwise convolution. Since the input and output feature mappings of the improved residual block maintain the same resolution and channel size, it is necessary to use the ordinary convolution block to extend the channel size by two times.

2.4. Spatial Channel Graph Convolutional Network

A graph convolutional network (GCN) extends the concept of convolution to graph-structured data by constructing adjacency matrices [29]. Initially applied in the field of knowledge graphs [30], GCN has more recently been employed for feature extraction from natural images. In a graph network, nodes receive potential information from their first-hop and nth-hop neighbors as messages propagate through the network. Information propagation in graph networks occurs along the edges between vertices in the graph. During graph convolution, each node in an input graph of dimensions C × W × H transforms its feature information and sends it to neighboring nodes to extract their feature information. Subsequently, each node aggregates feature information from its neighbors, integrating local structural information about the node. Finally, the aggregated information undergoes activation functions for nonlinear transformations to enhance the model's expression. Spatial channel graph convolutional network is illustrated in Figure 4.

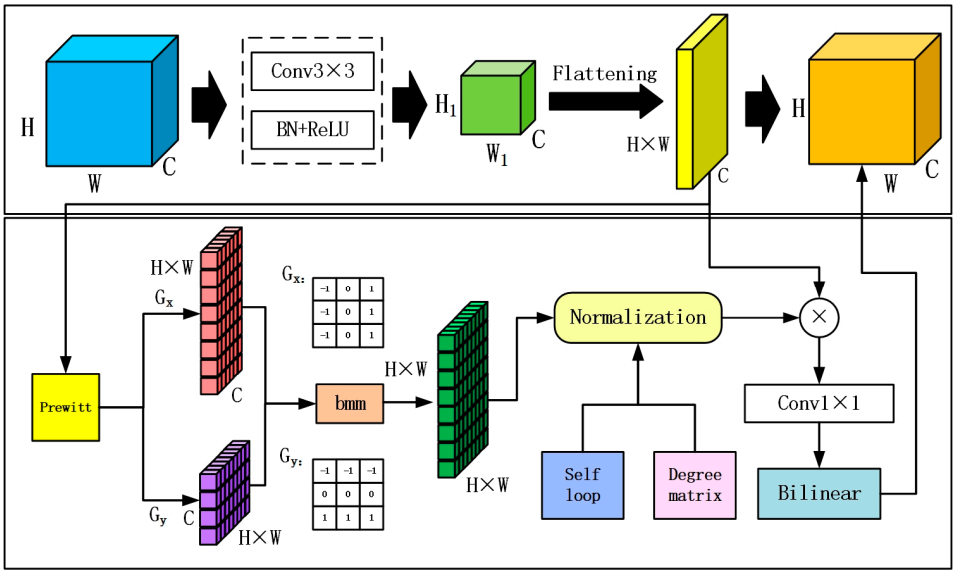


Figure 4. Spatial channel graph convolutional network diagram.

GCN is a type of neural network layer that creates a graph structure by selecting nodes and defining connections between them. In each layer, features are calculated using an adjacency matrix, which represents the connections between the nodes. The graph convolution at the l_{th} layer can be mathematically expressed as follows:

$$X^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}X^{(l)}W^{(l)}\right) \tag{4}$$

Here, σ denotes the nonlinear activation function, $\tilde{D}$ is the degree matrix of $\tilde{A}$, $\tilde{A}$ is the adjacency matrix, I_N is the identity matrix, l denotes the layer index in the GCN, $X^{(l)}$ represents the input/output features of all nodes in the l_{th} layer, and $W^{(l)}$ is the weight update matrix for the lth layer.

The Prewitt operator [31] is a type of differential operator utilized for identifying edges within images. It functions by analyzing the variances in pixel grayscale values within specific regions. In this research, the Prewitt operator is employed to construct an adjacency matrix by detecting changes in node values and providing precise gradient information. Here, G_x denotes the operator in the x-direction, which is used to calculate feature representations along the channel direction, while G_y represents the operator in the y-direction, which aids in spatial feature computation. The adjacency matrix is generated through matrix multiplication, with the formulation expressed as follows:

$$\tilde{A} = G_x\left(X_c{}^{(l)}\right)G_y\left(X_s{}^{(l)}\right) + I_N \tag{5}$$

In Equation (5), $G_x\left(X_c{}^{(l)}\right)$ and $G_y\left(X_s{}^{(l)}\right)$ represent the application of the Prewitt operator in the x and y directions, respectively, to detect gradients and create the adjacency matrix A. The product of these two matrices captures the gradient information across both spatial dimensions, and the identity matrix I_N is added to ensure self-loops, allowing for each node to retain its feature information. After constructing the adjacency matrix, the Laplacian symmetric normalization algorithm is applied to effectively preserve the structural information of the graph. This normalization process eliminates any imbalance in the adjacency matrix, promoting balanced relationships between nodes. Subsequently, the normalized adjacency matrix is multiplied with the original features to integrate the graph's topological structure information with node features. This fusion process enhances the capture of both local and global graph information. Following this, a series of 1D

convolutions and ReLU activation functions further abstract node features, making node representations more expressive. The final result is expressed as follows:

$$X^{(l+1)} = \sigma\left(F_c{}^{(l)} X^{(l)}\right) \tag{6}$$

where $F_c{}^{(l)}$ denotes the convolution.

The vertex information within the structure is propagated through three convolutional layers to aggregate information between individual features. Finally, interpolation operations are applied to restore the result to the original size C × W × H. The formulation is expressed as follows:

$$X^{(l+2)} = \sigma\left(F_c{}^{(l+1)} X^{(l+1)}\right) \tag{7}$$

$$X^{(l+3)} = \text{bilinear}\left(\sigma\left(F_c{}^{(l+2)} X^{(l+2)}\right)\right) \tag{8}$$

Here, bilinear represents bilinear interpolation.

2.5. scSE Module

The SE (squeeze-and-excitation) attention mechanism functions by learning feature weights based on loss values [32], thereby boosting the weights of effective feature maps and decreasing the weights of ineffective or less influential feature maps to optimize model performance. While incorporating the SE module into the original classification network may lead to increased parameters and computational complexity, the resultant performance improvement is generally deemed acceptable.

The fundamental concept of the scSE (spatial channel squeeze-and-excitation) module [33] involves the simultaneous operation of two sub-modules, cSE (channel squeeze-and-excitation) and sSE (spatial squeeze-and-excitation). This concurrent integration enhances information across both spatial and channel dimensions. Let the input feature map be denoted as U. The scSE structure is shown in Figure 5. The formulation is expressed as follows:

$$U = \left\{u^{(1,1)}, u^{(1,2)}, \ldots, u^{(i,j)}, \ldots, u^{(H,W)}\right\} \tag{9}$$

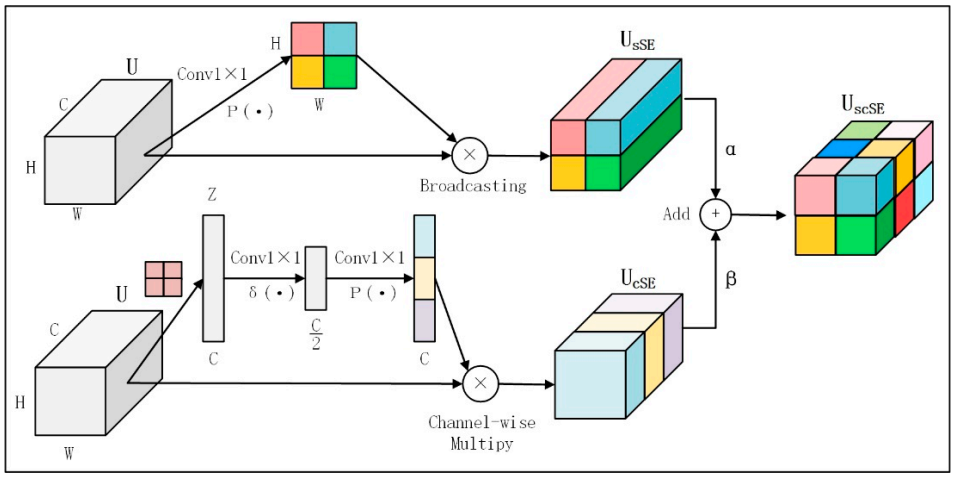


Figure 5. Squeeze-and-excitation attention mechanism diagram.

Here, $u^{(i,j)} \epsilon R^{C\times1\times1}$.

The cSE module integrates feature information across the channel dimension by using pooling and one-dimensional convolutions. It first reduces the dimensionality and then expands on attention weights to simplify the module complexity and improve computational speed. The formulation can be expressed as follows:

$$U_{cSE} = U \odot q \tag{10}$$

Here, q represents a weight tensor with the same number of channels as U, and $\odot$ denotes the element-wise multiplication operation.

The sSE module is similar in principle to the cSE module. It utilizes a convolution layer with a single channel and a 1×1 kernel size to process spatial feature information and derive spatial attention weights. These weights are subsequently normalized using the sigmoid function. The representation of the formula is as follows:

$$U_{sSE} = \left\{ \bigcup_{i=1}^{H} \bigcup_{j=1}^{W} P\left(*u^{(i,j)} \right) u^{(i,j)} \right\} \tag{11}$$

where $*$ denotes the convolution operation applied to the input feature map U, and P represents the sigmoid activation function, which normalizes the weights. The union operators U are used to concatenate the processed feature maps across both the height H and width W dimensions, forming the final spatial attention map.

To improve the module's grasp of contextual information, we introduce weights α and β, which are continually updated and learned during the training process. These weights are then utilized to carry out a weighted summation with the feature matrix, effectively balancing spatial and channel information. The formula for the output feature map U_{scSE} is as follows:

$$U_{scSE} = \alpha U_{sSE} + \beta U_{cSE} \tag{12}$$

3. Experimental Data and Evaluation Indexes

3.1. Experimental Data

To facilitate the successful implementation of railway transportation projects, it is essential to identify and address potential challenges, with a crucial task being the precise mapping of railway tracks. Currently, there is a relative scarcity of publicly available datasets for railway tracks, and existing datasets often do not meet the requirements for high-precision railway track segmentation in terms of clarity and reliability. To tackle this issue, this study autonomously developed a high-resolution railway track image dataset captured by drones, covering specific train station areas in southern China.

Due to the specific requirements of semantic segmentation models for input image resolution and format, raw image data cannot be directly used for model training. Therefore, a series of preprocessing operations are necessary, including image cropping, annotation, and format conversion, to align with the model's input requirements. Additionally, to improve the dataset's generalization and minimize the risk of overfitting, this study utilized data augmentation techniques to manually expand the original dataset. Operations such as cropping, rotation, and flipping of images effectively increased the diversity of the dataset, while removing redundant and duplicate data items. It is important to note that certain regions within the dataset do not contain railway tracks, such as rivers, fields, or houses, resulting in some images having no pixel points. To address this, a sliding window algorithm was applied to crop the dataset images, ensuring that images without track annotations did not negatively impact the extraction performance.

In summary, this study successfully compiled a railway track dataset comprising 13,285 images, focusing on specific areas of railway tracks. The dataset was meticulously divided into training and validation sets in a 9:1 ratio to support systematic model training and performance evaluation. Figure 6 showcases sample images from the constructed railway dataset in this study, providing rich and high-quality resources for railway track segmentation tasks.

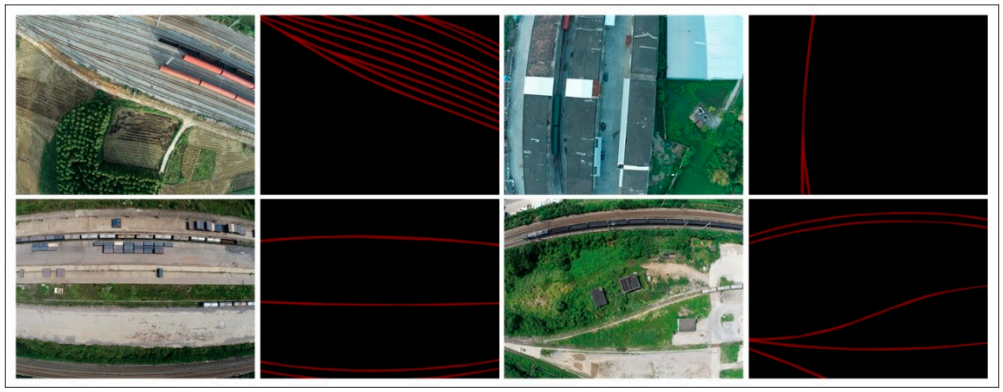


Figure 6. Original image and label image (railway dataset).

In this study, the DeepGlobe dataset [31] was used to provide enriched experimental results. This dataset comprises 3984 aerial images, each with dimensions of 1024 × 1024 pixels. To ensure compatibility with the model used, the labeled images in the dataset were converted to grayscale during preprocessing. Subsequently, the dataset was split into 3585 training images and 399 validation images. In the images, the background is represented in black, and the roads are represented in white. Figure 7 illustrates the images from the DeepGlobe dataset along with their corresponding labels.

Figure 7. Original image and label image (DeepGlobe dataset).

3.2. Experimental Environment

The experimental setup in this study was conducted on a Windows 11 system with an NVIDIA GeForce RTX 3060 Laptop GPU. Python version 3.9 and CUDA version 11.7 were used for experimentation. The training process utilized the Nadam optimizer, an extension of gradient descent optimization algorithms that combines Adam (adaptive moment estimation) and NAG (Nesterov accelerated gradient), thereby enhancing optimization performance. The initial learning rate was set to 0.001, with a weight decay coefficient of 1×10^{-4} to mitigate the risk of overfitting. Momentum was configured at 0.95, with a momentum decay coefficient of 0.004. The batch size was set to 4, and the model underwent 100 epochs of training.

3.3. Evaluation Index

In this pixel-based semantic segmentation model, several evaluation metrics are utilized, including Intersection over Union (IoU), Recall, Precision, and F1-Score. IoU measures the overlap between the predicted segmentation and the ground truth segmentation, divided by the union of the predicted and ground truth segments. Recall indicates how many positives in the sample are correctly predicted, Precision assesses the accuracy of positives predicted by the model, and F1-Score is the harmonic mean of Precision and Recall. The formulas for these evaluation metrics are as follows:

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{13}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{14}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{15}$$

$$\text{F1 Score} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{16}$$

Here, TP represents the number of pixels correctly extracted as railway tracks, FP denotes the number of pixels misclassified as railway tracks (when they are actually background), TN represents the number of pixels correctly classified as background, and FN indicates the number of pixels erroneously classified as background (when they are actually railway tracks).

3.4. Experimental Analysis

3.4.1. Visual Analysis of Loss Function

To visually show how our model performs, we examined the loss function convergence during training and validation on the railway and DeepGlobe datasets, as shown in Figures 8 and 9. Using the same experimental conditions with 100 iterations, all models displayed a gradual decrease in loss values over training epochs, almost converging by the 80th epoch.

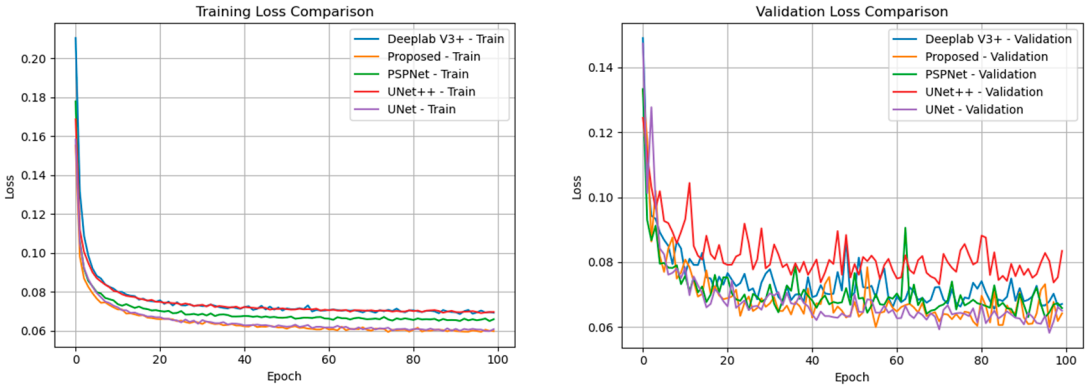


Figure 8. Railway dataset training and validation curves.

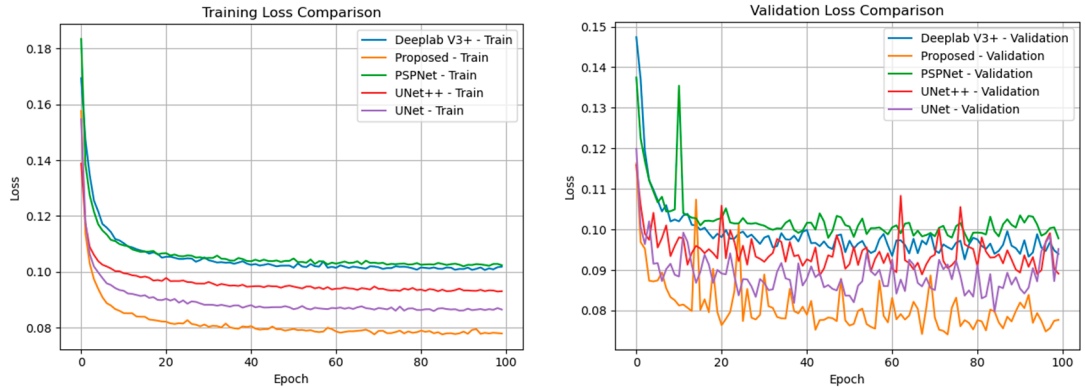


Figure 9. DeepGlobe dataset training and validation curves.

In the experiments conducted on the railway dataset, the proposed model demonstrated superior performance compared to others in terms of the speed of loss convergence and the final training loss. DeepLab V3+ and PSPNet showed rapid reduction in loss at the early stages, followed by quick stabilization with minimal fluctuation in validation loss. On the other hand, U-Net and UNet++ exhibited more volatile loss curves, indicating less

stable training and poorer generalization. In terms of validation, our model consistently achieved the lowest and most steady loss, demonstrating excellent generalization. DeepLab V3+ and PSPNet also achieved swift stabilization of validation loss after initial fluctuations, maintaining low levels throughout. In contrast, UNet++ and U-Net struggled with higher and erratic validation losses.

In the experiment conducted using the DeepGlobe dataset, all models exhibit a rapid decrease in loss during the initial epochs on the training set, which suggests effective learning. The proposed model demonstrates the best performance, as it has the lowest training loss. Additionally, UNet++ and UNet also perform well, with relatively low losses. On the other hand, DeepLab V3+ and PSPNet exhibit higher training losses, with PSPNet performing the worst.

When considering the validation set, the proposed model consistently maintains the lowest validation loss, indicating its excellent ability to generalize to unseen data. UNet and UNet++ also perform well on the validation set, although their losses are slightly higher. In contrast, DeepLab V3+ and PSPNet show higher and more fluctuating validation losses, suggesting poorer stability and weaker generalization ability.

3.4.2. Impact of the scSE on Model Performance

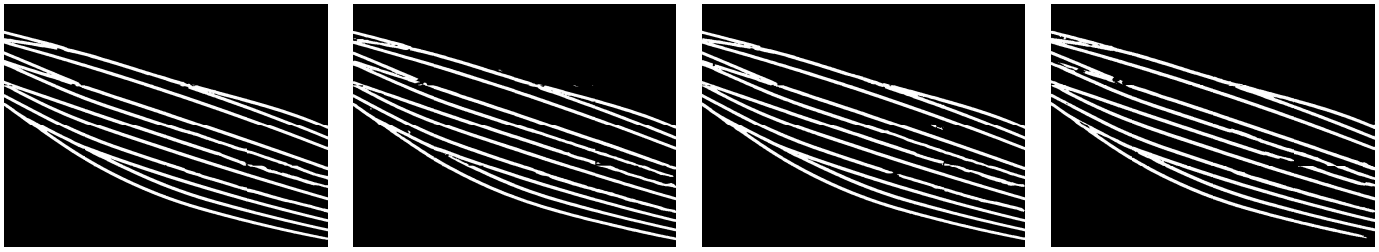
To achieve optimal performance of the model, we defined four separate positions for integrating scSE block. Subsequently, we conducted a comprehensive quantitative analysis to assess the optimal configuration. This analysis included various performance metrics such as MIoU, OA, and Recall. Furthermore, we generated and analyzed figures of the extraction results to visually compare the impact of each scSE block location on performance. Through this thorough evaluation, we identified the most effective scSE block locations, leading to enhanced overall performance and robustness of the model.

As demonstrated in Table 1, it can be observed that the scSE block significantly improves the segmentation quality at every position of the network. This enhancement is more noticeable when the scSE block is added after the decoder rather than the encoder. Moreover, when both positional configurations are combined, there is a more significant improvement in the overall model. These findings indicate that the optimal approach is to incorporate the scSE block after both the encoder and the decoder.

Table 1. Quantitative statistics depending on the position of the scSE.

Position	Railway			DeepGlobe		
	MIoU	**OA**	**Recall**	**MIoU**	**OA**	**Recall**
None	87.64%	96.25%	85.12%	66.54%	95.71%	79.65%
Only Encoder	87.23%	96.02%	87.14%	72.17%	95.94%	79.76%
Only Decoder	87.58%	97.37%	86.78%	72.39%	95.83%	79.88%
Both	88.47%	97.62%	88.90%	75.81%	96.02%	80.28%

As shown in Figure 10, the addition of the scSE block at different points in the network greatly affects the consistency and clarity of railway track segmentation. When the scSE blocks are used after both the encoder and decoder (configuration a), the results are the most coherent, with clear and continuous rail lines that are free from noise. This indicates that incorporating scSE blocks effectively enhances spatial and channel attention, thus preserving fine structural details. On the other hand, configuration (d) shows the poorest performance, with multiple breaks and uneven edges. This decline is due to the lack of attention mechanisms necessary for maintaining the integrity of the rail lines. Configurations (b) and (c) produce intermediate results, with fewer breaks but more noise. This suggests that attention at a single stage is not enough to fully capture and preserve the important details of the rail lines.



(**a**) (**b**) (**c**) (**d**)

Figure 10. Rail extraction results depending on different positions of the scSE. (**a**) Both Encoder and Decoder, (**b**) Only Encoder, (**c**) Only Decoder, (**d**) None.

When comparing the road extraction results to the ground truth (a), configuration (b) stands out by incorporating scSE blocks after both the encoder and decoder, resulting in superior performance. This configuration excels in capturing complex road topologies with minimal errors, demonstrating an optimal balance of spatial and channel information to enhance segmentation accuracy. On the other hand, configuration (c) produces more complete extractions but introduces a higher number of false positives, indicating that over-reliance on spatial information alone may lead to over-segmentation in certain areas. Configurations (d) and (e) fall short by missing several road segments, but they exhibit fewer artifacts, likely due to reduced network complexity. This reduction in complexity limits overfitting but also constrains the model's ability to capture intricate structural details. Road extraction results are shown in Figure 11.

(**a**) (**b**) (**c**) (**d**) (**e**)

Figure 11. Road extraction results depending on different positions of the scSE: (**a**) Ground Truth, (**b**) Both Encoder and Decoder, (**c**) Only Encoder, (**d**) Only Decoder, (**e**) None.

Each of the cSE and sSE blocks possesses unique properties, which contribute distinct advantages to the model. In our investigation to determine the optimal aggregation strategy of sSE and cSE, we explored five different possibilities. Among these, weighted fusion emerged as the best-performing method. The superior performance of weighted fusion can be attributed to its ability to assign optimal importance to both spatial and channel-wise features, thereby enhancing the model's ability to capture relevant information from the input data.

In comparison, the max-out and addition methods, while still effective, yielded slightly inferior results. The max-out method, which selects the maximum value from either the sSE or cSE outputs, can sometimes discard useful information by focusing solely on the most prominent features. Similarly, the addition method, which sums the outputs of the sSE and cSE blocks, may fail to appropriately balance the contribution of each type of feature, potentially leading to suboptimal feature representation. These factors contribute to their slightly lower performance compared to the more nuanced and adaptive weighted fusion strategy.

The multiplication and concatenation methods performed the worst, with all metrics lagging behind the others. Multiplication can overly emphasize some features while suppressing others, leading to potential information loss. Concatenation, by simply merging

the outputs, may introduce redundancy and dilute feature effectiveness. These factors contribute to their inferior performance compared to more effective strategies like weighted fusion. The comparison results of different aggregation strategies are presented in Table 2.

Table 2. Quantitative statistics depending on the aggregation strategies of the scSE.

Aggregation Strategies	Railway			DeepGlobe		
	MIoU	OA	Recall	MIoU	OA	Recall
Max-out	87.20%	95.56%	86.35%	77.41%	94.72%	79.43%
Addition	86.39%	95.08%	83.63%	77.63%	92.97%	77.45%
Multiplication	82.12%	94.10%	70.22%	70.61%	87.42%	75.34%
Concatenation	83.91%	93.72%	82.79%	76.23%	92.35%	78.29%
Weighted fusion	88.72%	96.49%	86.68%	78.49%	95.16%	80.26%

3.4.3. Comparative Experimental Analysis

To thoroughly evaluate our model's performance in railway track segmentation tasks, we selected several representative network architectures as controls, including U-Net, DeeplabV3+, PSPNet, and UNet++. We conducted experiments under consistent conditions and dataset settings, and the results and performance evaluations of our model and other network structures are summarized in Tables 3 and 4. Red frames indicate regions with large differences in extraction results. By comparing the detailed features of each model, we observed significant differences in their ability to identify railway track switches and sections obstructed by train cars.

Despite U-Net's widespread use in medical imaging segmentation, it performed poorly in track extraction tasks. It showed severe deficiencies in capturing track edge information and led to numerous discontinuous regions in its output. This resulted in a considerable deviation from ground truth annotations and highlighted its limitations in dealing with the complex, elongated structures typical of railway tracks. U-Net's inability to maintain track continuity resulted in fragmented segmentations, which are particularly detrimental in applications requiring high precision and reliability.

Table 3. Railway dataset model comparison table.

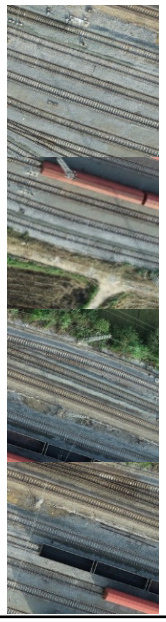
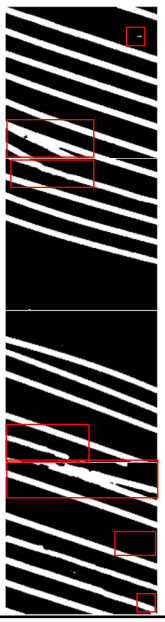
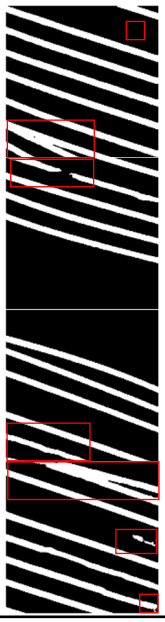
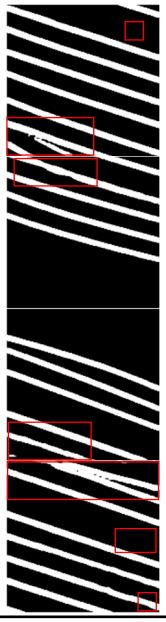
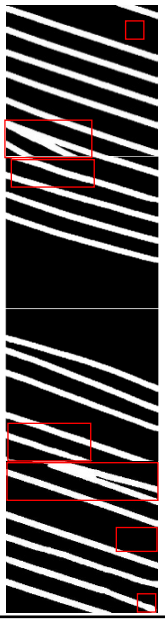
Original	PSPNet	UNet++	U-Net	Deeplab V3+	Proposed Method	Ground Truth
						

Table 4. Railway dataset quantitative statistics table.

Model	Rail IoU	MIoU	OA	F1 Score	Recall	Precision
PSPNet	77.89%	86.43%	96.24%	87.59%	87.65%	85.54%
U-Net	77.54%	86.24%	96.43%	87.35%	86.62%	85.19%
UNet++	76.72%	85.73%	95.67%	84.48%	87.71%	82.93%
Deeplab V3+	76.55%	86.03%	97.08%	87.14%	88.45%	85.88%
Propose Method	**79.49%**	**87.98%**	**97.48%**	**88.18%**	**89.79%**	**88.57%**

UNet++, an extension of U-Net, showed improvements over its predecessor in certain aspects but still exhibited extraction omissions in local areas, particularly under conditions where train cars caused occlusions. These occlusions led to noticeable discontinuities in the segmentation results, undermining the model's effectiveness. Despite its architectural enhancements, UNet++ struggled to handle the occlusion challenges inherent in railway environments, reflecting its limitations in real-world applications.

PSPNet and DeeplabV3+ demonstrated better performance in some respects. PSPNet, with its pyramid pooling module, is adept at capturing context at multiple scales, and DeeplabV3+, with its atrous spatial pyramid pooling, excels at maintaining resolution and capturing fine details. However, both models still suffered from multiple missed detections during identification tasks. The segmented outputs often presented jagged track edges, lacking the smoothness required for accurate track delineation. Additionally, these models struggled with the incomplete extraction of junction areas, which are critical for effective railway track management.

In the areas highlighted by red boxes, our model's segmentation results displayed notable advantages compared to other algorithms. Our model not only showed superior continuity and completeness in railway track identification but also demonstrated significantly improved smoothness and clarity of edges in contrast to comparative algorithms. While other models exhibited discontinuities and extraneous speckling in the extracted railway tracks, our model produced smooth edges with no breaks, demonstrating superior performance in maintaining the integrity of the track lines. The edge details were remarkably precise, with no significant deviations from the ground truth, ensuring reliable performance even in complex scenarios involving occlusions and track switches.

Our model's architecture benefits from enhanced feature extraction capabilities, possibly due to the incorporation of advanced attention mechanisms or multi-scale feature integration. These enhancements enable our model to maintain continuity and accurately segment occluded regions, outperforming other models in crucial evaluation metrics. The ability to handle fine-grained details and maintain high fidelity in edge representation is essential for practical applications in railway track maintenance and safety monitoring.

In summary, our model displayed high accuracy and robustness in railway track segmentation tasks, especially in handling track switches and edge details. It consistently outperformed other models in performance evaluations, showcasing its potential as a superior tool for railway infrastructure analysis. The advancements in our model's architecture address the limitations observed in other state-of-the-art networks, providing a dependable solution for detailed and accurate railway track segmentation.

The railway track segmentation dataset features high-resolution characteristics, clear image details, and consistent track continuity, which enable precise capture of track information. Upon quantitative analysis of experimental results, it is evident that the model proposed in this study outperforms other reference models across all evaluation metrics. Notably, the Intersection over Union (IoU) for railway tracks increased by 2% to 3% compared to other networks, accompanied by slight yet significant improvements in overall accuracy (OA), F1 score, and precision metrics. Particularly remarkable are the achieved high levels of recall (89.79%) and precision (88.57%).

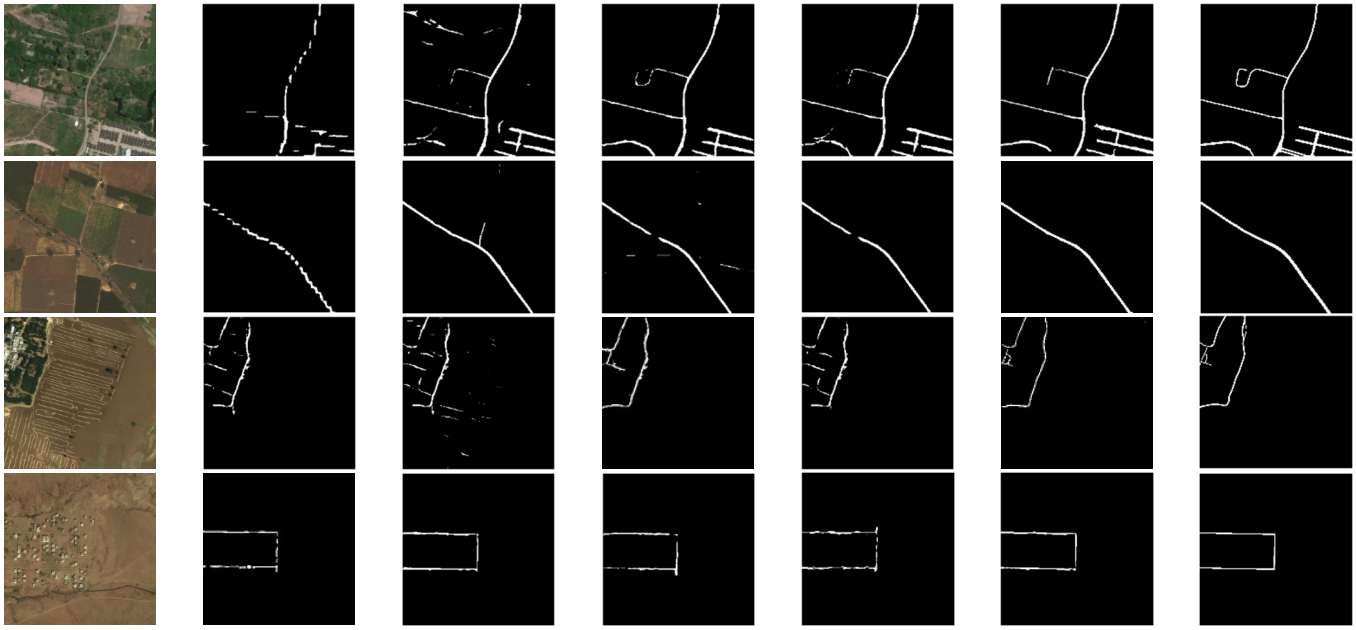
These substantial quantitative enhancements unequivocally validate the effectiveness of this model in improving the accuracy of railway track segmentation. Through optimized network architecture and algorithm design, this model has significantly improved its capa-

bility to recognize detailed features of railway tracks, especially in handling track switches and edge details, showcasing exceptional segmentation performance. Furthermore, the model's excellent performance in recall and precision further substantiates its reliability and accuracy in railway track segmentation tasks.

In order to comprehensively evaluate the model's ability to generalize, this study expanded its experimental scope to include the DeepGlobe dataset, which is known for its diverse background types. The DeepGlobe dataset presents certain differences in annotation accuracy compared to our custom-built railway dataset. While occlusion issues are relatively limited in the DeepGlobe dataset, the complexity of road structures, intricate road shapes, and background textures that resemble road colors pose significant challenges for segmentation tasks.

The comparative analysis presented in Table 5 demonstrates that the model proposed in this study excels in extraction performance, effectively minimizing missing parts and preserving topological integrity. This is particularly important in scenarios where maintaining the continuity and coherence of segmented objects is crucial. The robustness of our model against the varied and complex backgrounds in the DeepGlobe dataset underscores its capability to generalize well beyond the specific conditions of the railway dataset.

Table 5. DeepGlobe dataset model comparison table.

Original	PSPNet	UNet++	U-Net	Deeplab V3+	Proposed Method	Ground Truth
						

The results presented in Table 6 show that our model performs well across various evaluation metrics such as precision, recall, F1-score, and Intersection over Union (IoU). Although the Road IoU and Overall Accuracy (OA) of our model are 0.2% lower than those of Deeplabv3+ and UNet respectively, it exhibits a significantly higher recall rate, indicating its strong performance in capturing relevant features and minimizing false negatives.

In summary, the results demonstrate that our model not only performs well in diverse and challenging conditions but also maintains high accuracy and integrity in segmentation tasks. Its ability to handle intricate details and complex textures, while preserving the overall structure and continuity of segmented objects, highlights its superior generalization capabilities. These findings confirm the robustness and versatility of our model, making it a reliable choice for various segmentation applications across different datasets.

Table 6. DeepGlobe dataset quantitative statistics table.

Model	Road IoU	MIoU	OA	F1 Score	Recall	Precision
PSPNet	73.38%	69.80%	96.08%	77.90%	64.63%	76.56%
U-Net	73.07%	70.02%	**97.10%**	70.73%	76.22%	77.28%
UNet++	73.49%	75.00%	96.72%	78.11%	76.79%	75.03%
Deeplab V3+	**78.19%**	72.31%	96.49%	73.98%	66.68%	77.05%
Propose Method	77.96%	**77.33%**	96.98%	**79.00%**	**82.38%**	**78.13%**

3.4.4. Analysis of the Ablation Experiment

In order to further understand the roles of the graph convolutional network (GCN) module, residual chain module, and scSE attention mechanism module in railway track segmentation tasks, this study conducted ablation experiments on both the railway track dataset and the DeepGlobe dataset. The $\surd$ and $\times$ in the table indicate whether the module is added or not, respectively.

The results of the ablation experiments on the railway track segmentation dataset indicate that the standalone GCN module outperforms the standalone residual chain module in terms of segmentation effectiveness. When combined, they show a slight performance improvement due to their synergistic interaction. Furthermore, the integration of the scSE attention mechanism module into the residual chain module significantly enhances precision, while its impact on the improvement of the GCN module's performance is relatively limited.

When these three modules work together, the overall performance metrics, including F1 score and precision, show significant improvements, while recall also reaches optimal levels. These findings offer quantitative evidence for understanding the individual roles of each module in railway track segmentation, which are summarized in Table 7.

Table 7. Railway dataset ablation experiment.

GCN	Residual	scSE	MIoU	F1 Score	Recall	Precision
$\surd$	$\times$	$\times$	87.93%	87.85%	89.42%	84.33%
$\times$	$\surd$	$\times$	86.90%	86.92%	89.40%	82.05%
$\surd$	$\surd$	$\times$	88.20%	88.46%	88.59%	85.35%
$\times$	$\surd$	$\surd$	88.36%	88.58%	89.44%	88.02%
$\surd$	$\times$	$\surd$	88.05%	86.98%	89.96%	84.59%
$\surd$	$\surd$	$\surd$	88.23%	89.25%	90.79%	89.62%

These findings further validate the effectiveness of the multi-module fusion strategy in enhancing segmentation task performance, especially in handling the complex backgrounds and intricate road structures present in the DeepGlobe dataset. The GCN module leverages its capability to capture topological structure information from images, providing additional advantages for segmentation tasks. The residual chain module enhances the model's ability to capture details by increasing network depth without increasing the computational burden. The scSE attention mechanism further enhances recognition accuracy by reinforcing important parts of the feature maps. The comparison results of the ablation experiments on the DeepGlobe dataset are presented in Table 8.

Table 8. DeepGlobe dataset ablation experiment.

GCN	Residual	scSE	MIoU	F1 Score	Recall	Precision
$\surd$	$\times$	$\times$	69.81%	76.77%	71.03%	76.23%
$\times$	$\surd$	$\times$	63.91%	77.25%	79.27%	76.09%
$\surd$	$\surd$	$\times$	67.01%	79.36%	79.85%	75.72%
$\times$	$\surd$	$\surd$	67.03%	79.84%	73.67%	77.66%
$\surd$	$\times$	$\surd$	70.00%	79.32%	73.26%	79.84%
$\surd$	$\surd$	$\surd$	76.45%	81.93%	81.86%	80.13%

4. Discussion

The extraction of railways from high-resolution images presents several challenges, including interference from train carriages, roadside buildings, shadows, complex environments, crisscrossing tracks, and varying weather conditions. Traditional convolutional neural network (CNN) architectures often struggle to accurately extract railways in such complex scenarios.

This study evaluates the performance of our proposed model through comparative experiments, ablation studies, and visual analysis of loss functions using two distinct datasets. The experimental results demonstrate that our model significantly improves the accuracy of railway extraction while effectively addressing occlusion and switch recognition issues.

To address occlusion, the model incorporates a stacked design of multi-scale convolutional layers, which expands the network's receptive field and enhances its ability to identify occluded regions. This design allows for the model to maintain a high level of detail in the presence of obstructing objects like train carriages and buildings. The improvement in occlusion handling not only leads to more precise segmentation of railway tracks but also minimizes the occurrence of false positives and negatives, which are common in complex environments.

Furthermore, the integration of a graph convolutional network (GCN) with the Prewitt operator notably enhances the recognition of intricate switch edges, further optimizing the extraction of occluded railways. Moreover, the combination of GCN with spatial attention mechanisms effectively aggregates feature information of topological structures to comprehensively extract complex switch configurations. This approach allows for the model to focus on relevant spatial features while ignoring irrelevant background information, thereby improving the overall extraction accuracy and robustness across different environmental conditions.

The introduction of depth-wise separable convolutions and residual connections reduces model complexity and parameter count, while optimizations in the decoder part, including activation functions and network depth, significantly contribute to improving railway extraction accuracy. These architectural choices enhance the model's efficiency, making it more suitable for real-time applications where computational resources are limited. Moreover, the reduced complexity does not compromise the model's performance, demonstrating that our design achieves an optimal balance between accuracy and efficiency.

The self-built railway dataset, which comprises extensive data volume and high-quality images showcasing various railway conditions and environmental features, enhances the credibility of experimental analyses on this dataset. The diversity of the dataset ensures that the model is trained on a wide range of scenarios, improving its generalization capability. Experimental results on the DeepGlobe public dataset further validate the model's generalization capability. Despite slight discrepancies in road extraction results compared to ground truth, our model demonstrates advantages over comparative models. The model's ability to generalize well across different datasets suggests that it can be effectively applied to a variety of railway and road extraction tasks beyond the specific conditions of our dataset. This versatility is a key strength of our approach.

Additionally, the utilization of cross-entropy loss balances differences between different data categories and updates model parameters through backpropagation, enabling the model to effectively extract railway and road information. This loss function ensures that the model does not overfit while effectively preventing gradient explosion from occurring, allowing for more balanced and accurate predictions across all categories.

5. Conclusions

This study presents a novel algorithm for segmenting railway tracks from high-resolution images. The algorithm integrates graph convolution with residual structures to address challenges such as incomplete topological structures, unclear edge details, and interference from train carriages. It employs a graph convolutional network (GCN) in combination with a residual learning framework within an encoder–decoder architecture

to effectively merge graph structures and convolution operations, to enhance accuracy by capturing long-range dependencies and spatial information. Additionally, spatial squeeze-and-excitation (scSE) modules are incorporated to improve the model's capacity to capture complex railway features and contextual information, thereby enhancing segmentation accuracy.

The algorithm underwent extensive testing using high-definition aerial data from railway stations and the Massachusetts dataset. It was then compared to existing techniques. The results showed that the algorithm outperformed other methods in segmenting railway tracks. This research not only introduces an innovative solution to the problem of railway track segmentation in high-resolution images but also suggests new directions for further exploration in related fields.

Despite achieving significant improvements in segmentation accuracy, the algorithm has limitations. It should be enhanced to process images under extreme weather conditions and reduce its computational complexity for deployment on resource-constrained devices. Future research could focus on optimizing the algorithm and exploring different types of attention mechanisms to comprehensively capture features of railways and their surrounding environments, thereby further enhancing its performance.

Author Contributions: Conceptualization, Yanbin Weng and Meng Xu; railway image acquisition, Yanbin Weng and Xiahu Chen; data preprocessing, Meng Xu, Hua Yin and Hui Xiang; formal analysis, Meng Xu, Cheng Peng and Peixin Xie; funding acquisition, Yanbin Weng, Xiahu Chen and Cheng Peng; writing—original draft preparation, Yanbin Weng and Meng Xu; experiment, Meng Xu; writing—review and editing, Yanbin Weng, Meng Xu, Cheng Peng and Hui Xiang; paper revision, Yanbin Weng, Meng Xu and Cheng Peng. All authors have read and agreed to the published version of the manuscript.

Funding: This work was financially supported by the National Key Research and Development Program of China (2021YFF0501101), the National Natural Science Foundation of China (52172403, 62373178), the project of Hunan Provincial Department of Education of China (22B0577), and Natural Science Foundation of Hunan Province of China (2024JJ7139).

Data Availability Statement: Public datasets can be found at http://deepglobe.org/index.html (accessed on 20 October 2022). Other data in this study can be requested by contacting the corresponding author.

Acknowledgments: Firstly, the authors would like to express their gratitude for the financial support provided by the National Key Research and Development Program of China, the National Natural Science Foundation of China, and the Hunan Provincial Natural Science Foundation. Secondly, the authors would like to thank Zhuzhou Taichang Electronic Information Technology Co., Ltd. for their invaluable data and technical assistance.

Conflicts of Interest: Author Xiahu Chen was employed by the company Zhuzhou Taichang Electronic Information Technology Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Shvetsov, A.V.; Dronichev, A.V.; Kuzmina, N.A.; Shvetsova, S.V. Analysis of the Directions of Optimization of the Process of Ensuring Transportation Security in Railway Transport. *Transp. Res. Procedia* **2023**, *68*, 579–584. [CrossRef]
2. Gholamizadeh, K.; Zarei, E.; Yazdi, M. Railway Transport and Its Role in the Supply Chains: Overview, Concerns, and Future Direction. In *The Palgrave Handbook of Supply Chain Management*; Sarkis, J., Ed.; Springer International Publishing: Cham, Switzerland, 2024; pp. 769–796. ISBN 978-3-031-19883-0.
3. Song, T.; Schonfeld, P.; Pu, H. A Review of Alignment Optimization Research for Roads, Railways and Rail Transit Lines. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 4738–4757. [CrossRef]
4. Chen, S.-B.; Ji, Y.-X.; Tang, J.; Luo, B.; Wang, W.-Q.; Lv, K. DBRANet: Road Extraction by Dual-Branch Encoder and Regional Attention Decoder. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 3002905. [CrossRef]
5. Mo, S.; Shi, Y.; Yuan, Q.; Li, M. A Survey of Deep Learning Road Extraction Algorithms Using High-Resolution Remote Sensing Images. *Sensors* **2024**, *24*, 1708. [CrossRef] [PubMed]

6. Chen, H.; Li, Z.; Wu, J.; Xiong, W.; Du, C. SemiRoadExNet: A Semi-Supervised Network for Road Extraction from Remote Sensing Imagery via Adversarial Learning. *ISPRS J. Photogramm. Remote Sens.* **2023**, *198*, 169–183. [CrossRef]
7. Xu, H.; He, H.; Zhang, Y.; Ma, L.; Li, J. A Comparative Study of Loss Functions for Road Segmentation in Remotely Sensed Road Datasets. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *116*, 103159. [CrossRef]
8. Khan, N.; Shahid, Z.; Alam, M.M.; Sajak, A.A.B.; Nazar, M.; Mazliham, M.S. A Novel Deep Learning Technique to Detect Electricity Theft in Smart Grids Using AlexNet. *IET Renew. Power Gener.* **2024**, *18*, 941–958. [CrossRef]
9. Qian, L.; Huang, H.; Xia, X.; Li, Y.; Zhou, X. Automatic Segmentation Method Using FCN with Multi-Scale Dilated Convolution for Medical Ultrasound Image. *Vis. Comput.* **2023**, *39*, 5953–5969. [CrossRef]
10. Ansari, M.Y.; Yang, Y.; Meher, P.K.; Dakua, S.P. Dense-PSP-UNet: A Neural Network for Fast Inference Liver Ultrasound Segmentation. *Comput. Biol. Med.* **2023**, *153*, 106478. [CrossRef]
11. Jabbar, A.; Naseem, S.; Mahmood, T.; Saba, T.; Alamri, F.S.; Rehman, A. Brain Tumor Detection and Multi-Grade Segmentation through Hybrid Caps-VGGNet Model. *IEEE Access* **2023**, *11*, 72518–72536. [CrossRef]
12. Wang, J.; Song, J.; Chen, M.; Yang, Z. Road Network Extraction: A Neural-Dynamic Framework Based on Deep Learning and a Finite State Machine. *Int. J. Remote Sens.* **2015**, *36*, 3144–3169. [CrossRef]
13. 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 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 182–186.
14. Kipf, T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv* **2016**, arXiv:1609.02907.
15. Lu, Y.; Chen, Y.; Zhao, D.; Chen, J. Graph-FCN for Image Semantic Segmentation. In *Advances in Neural Networks—ISNN 2019*; Lu, H., Tang, H., Wang, Z., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2019; Volume 11554, pp. 97–105. ISBN 978-3-030-22795-1.
16. Chen, L.; Zhang, Q. DDGCN: Graph Convolution Network Based on Direction and Distance for Point Cloud Learning. *Vis. Comput.* **2023**, *39*, 863–873. [CrossRef]
17. Yu, H.; Ai, T.; Yang, M.; Huang, L.; Gao, A. Automatic Segmentation of Parallel Drainage Patterns Supported by a Graph Convolution Neural Network. *Expert Syst. Appl.* **2023**, *211*, 118639. [CrossRef]
18. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. Mobilenetv2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520.
19. Dai, Y.; Li, C.; Su, X.; Liu, H.; Li, J. Multi-Scale Depthwise Separable Convolution for Semantic Segmentation in Street–Road Scenes. *Remote Sens.* **2023**, *15*, 2649. [CrossRef]
20. Huang, T.; Chen, J.; Jiang, L. DS-UNeXt: Depthwise Separable Convolution Network with Large Convolutional Kernel for Medical Image Segmentation. *Signal Image Video Process.* **2023**, *17*, 1775–1783. [CrossRef]
21. Zhang, Y.; Xu, S.; Zhang, L.; Jiang, W.; Alam, S.; Xue, D. Short-Term Multi-Step-Ahead Sector-Based Traffic Flow Prediction Based on the Attention-Enhanced Graph Convolutional LSTM Network (AGC-LSTM). *Neural Comput. Appl.* **2024**, 1–20. [CrossRef]
22. Zhu, J.-Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2223–2232.
23. Weng, Y.; Huang, X.; Chen, X.; He, J.; Li, Z.; Yi, H. Research on Railway Track Extraction Method Based on Edge Detection and Attention Mechanism. *IEEE Access* **2024**, *12*, 26550–26561. [CrossRef]
24. Weng, Y.; Li, Z.; Chen, X.; He, J.; Liu, F.; Huang, X.; Yang, H. A Railway Track Extraction Method Based on Improved DeepLabV3+. *Electronics* **2023**, *12*, 3500. [CrossRef]
25. Saeedizadeh, N.; Minaee, S.; Kafieh, R.; Yazdani, S.; Sonka, M. COVID TV-Unet: Segmenting COVID-19 Chest CT Images Using Connectivity Imposed Unet. *Comput. Methods Programs Biomed. Update* **2021**, *1*, 100007. [CrossRef]
26. Zhao, S.; Feng, Z.; Chen, L.; Li, G. DANet: A Semantic Segmentation Network for Remote Sensing of Roads Based on Dual-ASPP Structure. *Electronics* **2023**, *12*, 3243. [CrossRef]
27. Mehta, S.; Rastegari, M. MobileViT: Light-Weight, General-Purpose, and Mobile-Friendly Vision Transformer 2022. *arXiv* **2022**, arXiv:2110.02178.
28. 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]
29. Li, X.; Yang, Y.; Zhao, Q.; Shen, T.; Lin, Z.; Liu, H. Spatial Pyramid Based Graph Reasoning for Semantic Segmentation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 8950–8959.
30. Yu, D.; Yang, Y.; Zhang, R.; Wu, Y. Knowledge Embedding Based Graph Convolutional Network. In Proceedings of the WWW '21: The Web Conference 2021, Ljubljana, Slovenia, 19 April 2021; pp. 1619–1628.
31. Balochian, S.; Baloochian, H. Edge Detection on Noisy Images Using Prewitt Operator and Fractional Order Differentiation. *Multimed. Tools Appl.* **2022**, *81*, 9759–9770. [CrossRef]

32. Zhu, W.; Li, H.; Cheng, X.; Jiang, Y. A Multi-Task Road Feature Extraction Network with Grouped Convolution and Attention Mechanisms. *Sensors* **2023**, *23*, 8182. [CrossRef]
33. Roy, A.G.; Navab, N.; Wachinger, C. Concurrent Spatial and Channel 'Squeeze & Excitation' in Fully Convolutional Networks. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2018*; Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2018; Volume 11070, pp. 421–429. ISBN 978-3-030-00927-4.