remote sensing

MDPI

Article

U3-Road: A CNN and Transformer Multi-Level Nested Model for Road Extraction from Remote Sensing Images

Hengfei Zhan [1,2] and Yunpeng Wang [1,2,*]

[1] State Key Laboratory of Deep Earth Processes and Resources, Guangzhou Institute of Geochemistry, Chinese Academy of Sciences, Guangzhou 510640, China; zanhengfei20@mails.ucas.ac.cn
[2] College of Earth and Planetary Sciences, University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: wangyp@gig.ac.cn

Highlights

What are the main findings?

- A CNN-Transformer fused remote sensing road extraction model, U3-Road, is constructed. It adopts an encoder–decoder framework and a multi-path nested structure to efficiently capture multi-scale spatial and contextual features of road targets, and uses a composite loss optimization strategy to further enhance road boundary identification and suppress road breakage defects.
- U3-Road achieves superior extraction accuracy and topological continuity while maintaining low parameter counts and computational FLOPs, breaking the conventional trade-off between model precision and computational overhead in existing road extraction algorithms.

What are the implications of the main findings?

- Optimized architecture design and composite boundary constraint strategies provide a reliable paradigm for solving topological discontinuity problems in remote sensing road networks under complex ground-object interference.
- The high-efficiency, low-cost, and highly scalable U3-Road framework supports large-scale automatic road mapping and can be widely applied to urban digital construction, intelligent transportation infrastructure investigation, and resource-limited remote sensing engineering tasks.

Abstract

As a crucial prerequisite for intelligent transportation and urban digitalization, the importance of road extraction is evident. In recent years, the rapid development of remote sensing technology and deep learning has made road extraction from remote sensing images a research focus in both academia and industry. However, the topological complexity of road networks, interference from ground-object occlusion, and the spectral and spatial limitations of remote sensing images have presented numerous challenges to the design and optimization of related algorithms. To address these issues, this paper proposes a road extraction algorithm, U3-Road, which combines the advantages of CNN and Transformer architectures. This algorithm adopts an encoder–decoder structure and uses multi-path nesting technology in the encoder and bridge modules, enabling the extraction of multi-scale contextual features and semantic information, which are fully integrated with the decoder. Additionally, it employs a composite loss function to focus on difficult-to-identify sections and constrain problems such as road breaks and deformations, thereby achieving better accuracy and connectivity. Tests on the DeepGlobe, Massachusetts, and CHN6-CUG datasets show that the Intersection over Union (IoU) of U3-Road reaches 70.47%, 69.55%,

Check for updates

Academic Editor: Tania Stathaki

Received: 11 March 2026
Revised: 21 April 2026
Accepted: 5 May 2026
Published: 12 May 2026

Copyright: © 2026 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.

and 70.87%, demonstrating leading performance. It effectively alleviates the problems of road breaks and misidentification, and compared with other representative methods, U3-Road performs better in the completeness and accuracy of road extraction results, achieving the best overall extraction effect while maintaining a lower parameter count and computational cost. This model provides a new solution for road extraction from remote sensing images.

Keywords: remote sensing image; road extraction; multi-scale feature; CNN transformer; CNN; transformer

1. Introduction

Road information (referring to the geometric shape, topological structure, spatial position, connectivity relationship and scale characteristics of roads in remote sensing images) is a crucial fundamental geographic element, playing an irreplaceable role in both theoretical research and engineering practice [1]. Remote sensing images are collected over long distances by satellites and aircraft, offering economic efficiency, objectivity, and comprehensiveness. They enable broad information coverage and rapid acquisition across large spatial areas [2]. Unlike manual road mapping, which suffers from low efficiency, high cost, poor timeliness, limited working environments, and challenging data integration, extracting roads from remote sensing images meets the need for large-scale and dynamic road data. This provides essential decision-making support for urban intelligent transportation management, emergency rescue and disaster relief, and urban digital construction, etc. [3].

In recent years, remote sensing detection techniques and high-resolution image processing technologies have been continuously upgraded and iterated. The spatial resolution of remote sensing images has entered the sub-meter level, laying a comprehensive and abundant data foundation for road extraction tasks [4]. To tackle the technical challenges of accurate road extraction from remote sensing imagery, researchers worldwide have conducted in-depth investigations and extensive systematic studies in this field, proposing numerous technical solutions with both theoretical merit and practical value. Generally, these approaches can be divided into two categories: traditional methods and deep learning-based methods [5].

Conventional approaches for road extraction can be generally categorized into four types: threshold segmentation [6], edge detection and morphological processing [7], region growing [8], and template matching [9].

Threshold segmentation methods utilize the differences in grayscale and spectral features between roads and backgrounds to binarize remote sensing images via predefined thresholds, thereby distinguishing road regions from non-road areas. This method features a simple calculation pipeline and high computational efficiency, making it suitable for high-resolution remote sensing images with relatively uniform backgrounds. Nevertheless, it is highly vulnerable to interference from factors such as shadow occlusion, vehicle obstruction on road surfaces, and vegetation coverage, leading to unsatisfactory extraction performance in complex scenes [10].

Edge detection combined with morphological processing first locates road edges by identifying abrupt gray-level changes using edge detection algorithms, then connects broken road edges via morphological operations and suppresses noise interference to generate complete road contours. This approach can rapidly capture the linear characteristics of roads and performs well on images with distinct road edge contours. Nevertheless, it is

sensitive to disturbances such as shadows and textures, and tends to produce false edges, thereby reducing extraction accuracy [11].

The region growing method initiates from road seed points and gradually merges neighboring homogeneous pixels according to predefined growth criteria to construct a complete road region. This approach effectively preserves road connectivity and performs favorably in scenarios with relatively uniform road textures. However, its extraction results rely heavily on the selection of seed points, and under complex background conditions, the method is prone to erroneously merging non-road regions [12].

The template matching method requires the pre-construction of a road feature template. By sliding the template across the remote sensing image and computing the similarity between the template and the corresponding image regions, areas with similarity exceeding a predefined threshold are identified as roads. This method performs well in extracting regularly shaped roads, yet exhibits poor adaptability to road curvature and width variations, and its performance heavily relies on prior knowledge for template design [13].

Deep learning-based remote sensing road extraction methods can be mainly categorized into three types: CNN-based extraction methods, Transformer-based extraction methods, and CNN-Transformer hybrid architectures [14].

For CNN-based extraction methods, the core lies in leveraging the local receptive field properties of convolution kernels to capture spatial detail features in remote sensing images. The mainstream framework adopts an encoder–decoder architecture. The encoder reduces feature dimensions and extracts deep semantic information through successive convolution and downsampling operations, while the decoder gradually recovers the resolution of feature maps via upsampling and skip connections. This pipeline ultimately achieves pixel-level binary classification between roads and the background. Such methods can accurately characterize local features including road edges and pavement textures, and perform reliably in extracting small- and medium-scale roads. However, constrained by the local receptive mechanism of convolution kernels, they struggle to effectively model long-range dependencies between distant pixels. Consequently, when dealing with large-scale road networks and complex scenes, these methods tend to generate discontinuous road extractions [15].

Transformer-based road extraction methods are centered on the self-attention mechanism. By computing attention weights between arbitrary pixel pairs within an image, these methods overcome the local receptive field constraint of CNNs and effectively capture long-range pixel dependencies across the entire image. Specifically, remote sensing images are split into fixed-size patches and embedded into sequence features. After modeling global contextual information via multi-head attention modules, the features are remapped to the pixel space to accomplish road segmentation. The advantages of such methods include accurately capturing the continuity of large-scale road networks and effectively alleviating problems such as road discontinuities and missed detections in complex scenarios. Nevertheless, due to the high computational complexity introduced by self-attention, these methods demand substantial computational resources, and their capability in preserving fine local details (e.g., road edges and narrow roads) is inferior to that of CNN-based approaches [16].

Hybrid architectures represent the mainstream paradigm for remote sensing road extraction [17]. Their core idea is to integrate the strong local detail extraction ability of CNNs with the global dependency modeling capacity of Transformers, thereby achieving a favorable balance between accuracy and efficiency. Such architectures can preserve fine structural details such as road edges, narrow roads, and intersections, while maintaining the continuity of large-scale road networks. They mainly include two typical structures: serial and nested. The serial architecture first employs a CNN encoder to extract local

detailed features from the image, then feeds the deep features into a Transformer module to model global dependencies, and finally fuses multi-scale features via a decoder to output the segmentation result [18]. The nested architecture embeds Transformer modules within the CNN encoder or decoder network, enabling simultaneous local feature extraction and global correlation modeling, and improving the model's adaptability to complex scenes through interactive feature fusion [19].

Deep learning-based hybrid architectures have achieved remarkable progress in remote sensing road extraction. Nevertheless, three critical limitations still persist:

Such architectures must jointly conduct convolutional operations of CNNs and self-attention computations of Transformers. The integration of these two modules substantially increases the parameter count and computational burden, particularly for high-resolution images. Global dependency modeling using multi-head attention consumes considerable computational resources, leading to prolonged training time and slow inference speed, making it difficult to satisfy real-time demands such as emergency rescue tasks [20].

Current feature fusion strategies mostly rely on simple concatenation or weighted summation, which cannot fully exploit the complementary advantages of CNN-derived local detailed features and Transformer-based global semantic features. When extracting complex road structures such as intersections and narrow roads, feature redundancy or inconsistency frequently arises, causing blurred boundaries and missed detections of small road segments [21].

When confronted with complex disturbances such as shadow occlusion and vegetation coverage, as well as challenging landscapes including mountainous and rural areas, the model's robustness and feature discriminability remain inadequate. This easily leads to false predictions and large fluctuations in extraction accuracy [22].

In view of the aforementioned challenges, this study proposes an encoder–decoder network, U3-Road, which integrates the advantages of CNNs and Transformers for road extraction from remote sensing images. The core principles of U3-Road follow: (1) in the encoding and bridging stages, multiple feature extraction branches are constructed to fully capture global context information and multi-scale semantic features; (2) in the decoding stage, a lightweight and efficient feature fusion strategy is adopted to achieve precise aggregation of remote sensing road target features; and (3) a targeted loss function is designed to balance pixel-level classification accuracy and the continuity of road topology in the remote sensing road extraction task. The experimental results show that this model not only outperforms the benchmark model in terms of accuracy, but also maintains lower parameter quantity and computational cost. It achieves better balance between accuracy and efficiency, and thus has excellent practical engineering application value.

2. Proposed Model

This study proposes a deep learning network, U3-Road, based on an encoder–decoder architecture. It uses a nested fusion framework to combine the local detail representation capability of CNN and the global correlation modeling ability of Transformers, thereby improving detail extraction accuracy and road continuity in the road extraction task. The overall network architecture of U3-Road is shown in Figure 1.

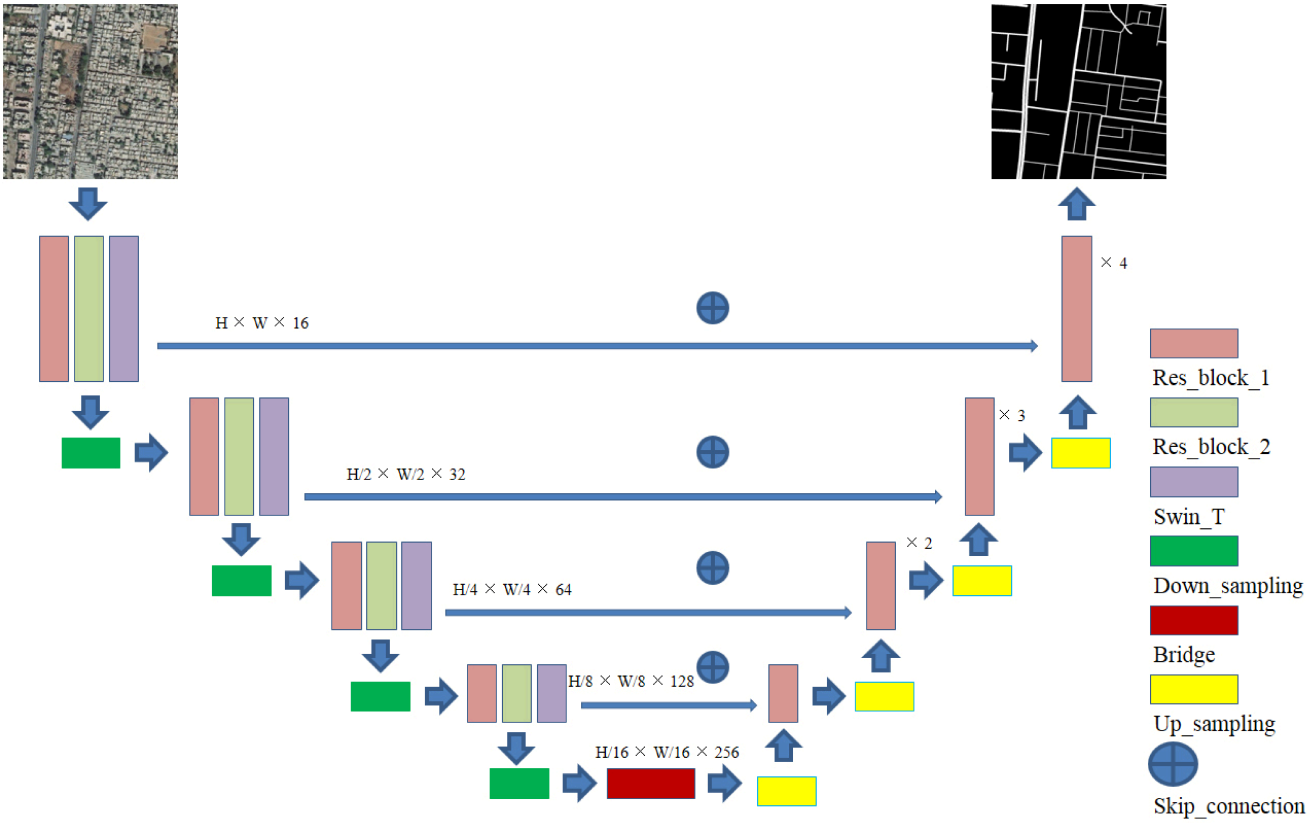

Figure 1. U3-Road network architecture diagram.

2.1. Encoder

The left part of U3-Road is the encoder, designed to extract multi-scale features from the image step by step. It compresses the spatial dimensions through downsampling and expands the receptive field, ultimately outputting abstract features rich in high-level semantics, which provide a contextual basis for the precise segmentation performed by the decoder. This part consists of four layers. Each layer uses the Res_block_1, Res_block_2, and Swin Transformer modules to extract road features. In addition to increasing depth, it also connects with the corresponding layers of the decoder. The network structures of the three modules are shown in Figure 2. The input and output dimensions of these modules are consistent, enabling seamless stacking and integration into various high-demand, fine-grained task scenarios. Through deep stacking, the model's performance limit is fully realized.

In the Res_block_1 module, a 3 × 3 standard convolution is combined with the residual network [23] to cover the entire neighborhood around each pixel. With each additional layer of 3 × 3 standard convolution, the receptive field size increases by 2. This systematic expansion ensures the capture of global context without losing pixel-level local details. U3-Road applies four 3 × 3 standard convolutions, expanding the receptive field to 9 × 9, which is used to match the detailed features of the road.

In the Res_block_2 module, a 3 × 3 dilated convolution [24] combined with the residual network is used. Without increasing the number of parameters or reducing the resolution of the feature map, the receptive field can be flexibly expanded. The calculation method for the receptive field (RFeq) of a single layer of dilated convolution is shown in Equation (1), and the receptive fields of the four different dilation rate sub-modules are shown in Figure 3. First, standard convolution fuses information around the pixels, and then the dilation rate is gradually increased to effectively prevent information loss during

the dilated convolution process. Finally, image information with a receptive field of 21 × 21 is obtained, which can bypass obstacles and maintain road continuity.

$$RF_{\text{eq}} = k + (k-1)(r-1) \quad (1)$$

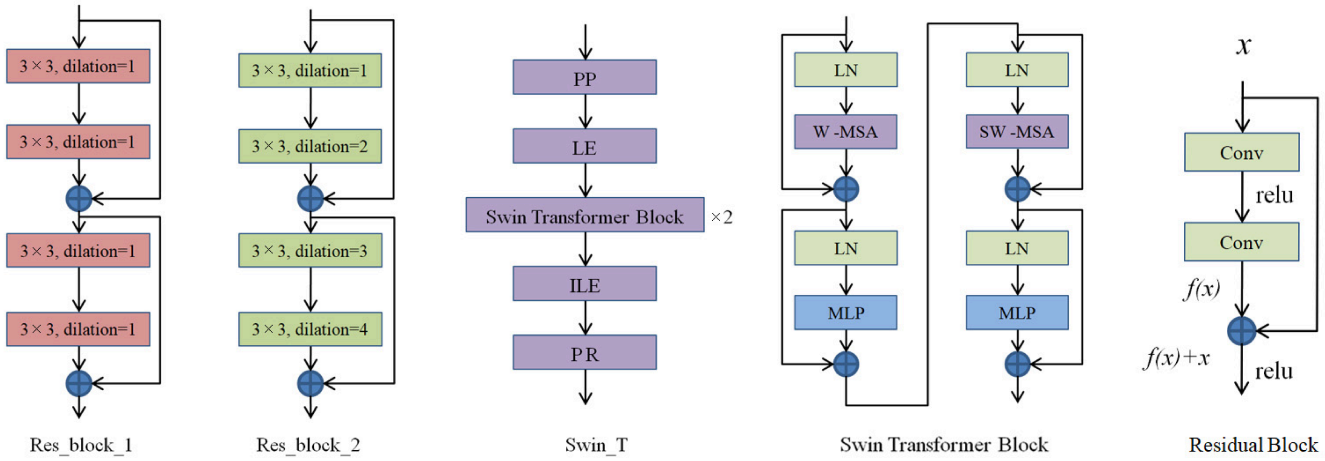

Figure 2. Structure of the core module of the model encoder. Res_block_1, Res_block_2, and Swin_T are the main modules of the encoder, which are used to extract road features from remote sensing images in parallel. Swin Transformer Block is the core unit of Swin_T, which is used for efficient global modeling and generating multi-scale features. Residual Block is a basic residual structure that exists in multiple modules mentioned above, which is used to alleviate gradient disappearance and enhance the feature extraction ability.

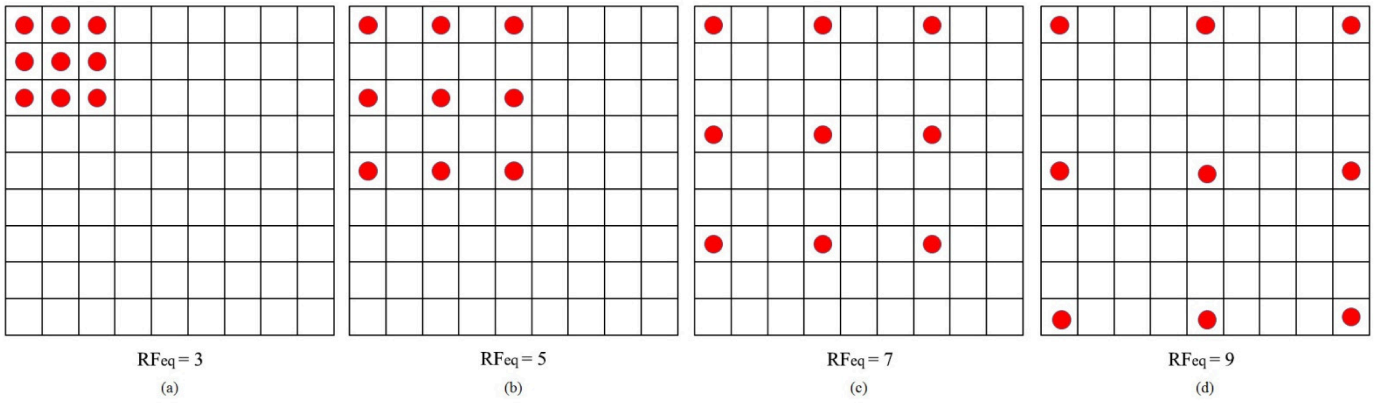

Figure 3. Comparison of receptive field ranges of ordinary convolution and dilated convolution with 3 × 3 convolution kernels: (**a**) ordinary convolution; (**b**) dilated convolution with a dilation rate of 2 (dilation = 2); (**c**) dilated convolution with a dilation rate of 3; (**d**) dilated convolution with a dilation rate of 4.

Here, k represents the original size of the convolution kernel, and r represents the dilation rate (also known as porosity, referring to the number of pixels between the sampling points in the dilated convolution, used to expand the receptive field).

The Swin_T module is a lightweight visual feature extraction unit based on the core architecture of the Swin Transformer [25]. It uses the PP (Patch Partition) sub-module to divide the input feature map into fixed-size, non-overlapping patches, mapping 2D spatial pixels to 1D sequence features. The LE (Linear Embedding) sub-module then transforms the dimensions of these local image features and aggregates information to generate token sequences suitable for Transformer modeling, while retaining effective feature information. The Swin Transformer Block restricts self-attention calculations to within each window, eliminating the need to model pixel correlations across the entire global feature map. This

reduces computational complexity from $O(n^2)$ to $O(n)$, significantly lowering computational overhead, and introduces the shifted window attention mechanism. By staggering adjacent hierarchical windows, it overcomes window boundary restrictions and realizes cross-window information interaction and feature fusion. This effectively expands the model's receptive field, ensuring the ability to capture local details while enhancing the modeling of global semantic information, and balancing the fine details of local features with the consistency of global features. The token sequence output is reconverted into feature maps using the ILE (Inverse Linear Embedding) and PR (Patch Reshape) sub-modules, facilitating integration with convolutional operations.

Through the encoder's three different modules, multi-scale context features are extracted and deeply fused to precisely focus on the main road areas in remote sensing images. By fully utilizing global and local semantic information, the approach aims to eliminate interference from complex backgrounds such as buildings, vegetation, and shadows, clarify road segments blocked by obstacles, and improve the completeness and accuracy of road extraction. Figure 4 shows the different focus ranges of the encoder's three modules.

Figure 4. Schematic diagram showing the focus ranges of the three modules of the encoder: (**a**) remote sensing image; (**b**–**d**) comparison diagrams of the information extraction ranges of the Res_block_1, Res_block_2, and SwinTransformer modules. The colored boxes denote the receptive field range of each module, and the receptive field ranges at each layer of the network may vary.

2.2. Bridge Module

The bridge module between the encoder and decoder serves as the central hub of the road extraction task. It receives high-level semantic features from the encoder, transforms and optimizes these features, and provides high-quality, highly adaptable input features for the decoder. After the encoder performs multiple layers of downsampling, the output features at the end are the highest-level features, containing the most abstract semantics and the largest receptive field. These features carry essential information such as target categories and global structures. The bridge module must aggregate and condense multi-scale encoder features, not only receiving features from the last encoder layer but also integrating multi-scale features from higher encoder layers to prevent information loss from relying on single-scale features.

In this study, residual convolution and the Transformer multi-head attention mechanism [26] are used in the bridge module to further compress redundant information, enhance core semantics such as road areas and target contusions, and form highly condensed feature representations. These provide semantic anchor points for the upsampling reconstruction in the decoder. The specific structure is shown in Figure 5.

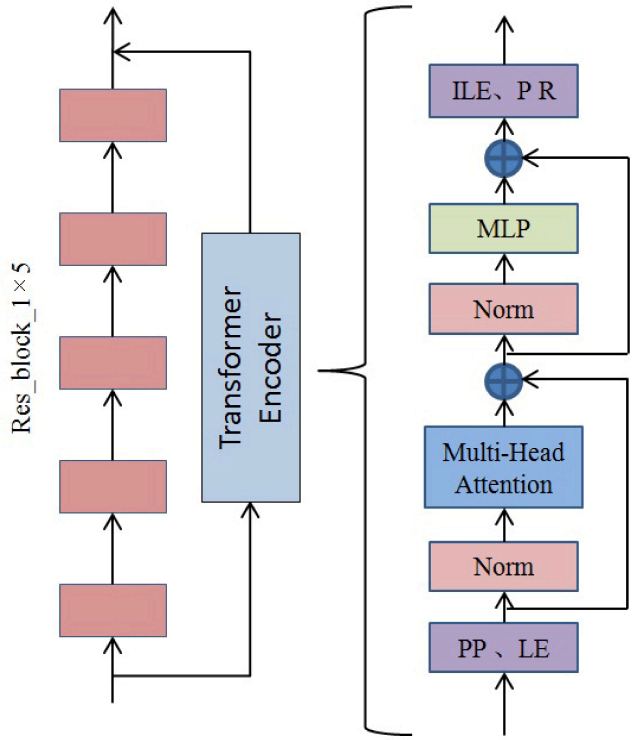

Figure 5. Bridge module structure diagram.

In deep networks, direct connections between the encoder and decoder are prone to issues such as information attenuation and gradient vanishing. The bridge module must optimize the transmission path for information and gradients, retain key information, and, through residual connections, preserve essential details from the encoder while compressing features. This approach avoids transmitting only abstract semantics and losing the fine information needed for decoder reconstruction. The residual mapping and short path connection structures of the bridge module allow gradients to flow directly from the decoder to the shallow layers of the encoder during backpropagation, alleviating the gradient-vanishing problem in deep networks and improving the stability and convergence speed of model training. As the final optimization step for encoder features, the bridge module can further improve feature quality, ensuring the decoder outputs high-precision results. By introducing the multi-head self-attention mechanism, road features are weighted and enhanced, enabling the decoder to focus more on key areas. With a modular architecture, the complexity can be adjusted according to task requirements. In high-performance computing scenarios, the module can be stacked multiple times to balance efficiency and performance.

2.3. Decoder

This study adopts a lightweight and efficient decoding architecture similar to LinkNet [27]. Following the design philosophy of Link-Net, the proposed decoder uses a streamlined encoder–decoder structure with symmetric skip connections to efficiently fuse encoder and decoder features. Compared with the decoders of U-Net [28] and Swin-UNet [29], it offers significant advantages in parameter count, computational demand, and inference efficiency.

The decoder unit primarily consists of the Res_block_1 module. The 3×3 standard convolution refines and corrects features, capturing local textures and edge information. The transposed convolution performs upsampling with a stride of 2, smoothing and restoring the spatial resolution of the feature map while avoiding feature distortion caused

by the checkerboard effect. The channel number is adjusted using a 1×1 convolution. This architecture features a highly modular design with a compact structure and no redundant computations. It facilitates engineering deployment and can be flexibly stacked for deep expansion, further improving model performance.

For features input through the encoder's skip connections, the decoder first achieves precise dimension alignment with a 1×1 convolution, and then uses element-wise addition to enhance the complementary semantics of the features. The encoder provides skip connections at three different levels of spatial semantic information, while the decoder features carry high-level global information. After fusion, this approach retains the detail and positioning accuracy of road targets and strengthens semantic consistency, significantly improving the completeness and accuracy of road extraction from remote sensing images. In contrast to the U-Net and SwinUNetdecoders, which use channel concatenation—doubling the number of channels and causing subsequent convolutional computation to grow exponentially—the element-wise addition fusion method does not change the feature channel dimension. This significantly reduces model parameters and computational cost, balancing performance and efficiency.

2.4. Loss Function

To balance pixel-level detection accuracy in road extraction from remote sensing images with the topological continuity of roads, this study constructs a composite loss function comprising Dice Loss [30], Focal Loss [31], and Skeleton Loss. This function quantifies the difference between the model's predictions and the true labels and serves as the optimization objective for backpropagation. During model training, the gradients of each network layer parameter with respect to the composite loss are computed using the chain rule, and parameter updates are performed through the gradient descent algorithm. Dice Loss, based on the pixel-level intersection-over-union ratio, effectively focuses on feature learning in the road target area and improves the overall recall rate of road pixels. Its calculation method is shown in Equation (2). Focal Loss introduces a difficulty-based sample weighting mechanism to reduce the loss contribution from easy-to-classify samples, such as background regions, forcing the model to focus on the feature fitting of difficult-to-classify areas like road edges and occluded sections. This approach alleviates topological breakage caused by prediction errors in challenging regions. The calculation process is shown in Equation (3). Skeleton Loss uses a weighted combination strategy to fuse skeleton pixel loss and skeleton length loss, specifically constraining pixel-level positioning accuracy and topological length consistency of the road centerline. This addresses issues such as road breakage, missed detection of narrow roads, and topological deformities that are difficult to constrain with pixel-level loss functions like cross-entropy. The specific calculation is shown in Equation (4).

$$L_{Dice} = 1 - \frac{2 \times \sum_i (p_i \times g_i) + \varepsilon}{\sum_i (p_i) + \sum_i (g_i) + \varepsilon} \tag{2}$$

where p_i represents the predicted value, g_i represents the label value, and ε is the minimum value, ensuring that the denominator is not zero.

$$L_{Focal} = -\alpha_t \times (1 - p_t)^{\gamma} \times log(p_t + \varepsilon) \tag{3}$$

where α_t represents the optional category weights, which are used to balance positive and negative samples. γ is the focusing parameter, which is used to suppress easily classified samples. p_t is the predicted probability that matches the label.

$$L_{Skeleton} = 1 - \frac{2 \times \sum_i \left(p_{Skel_i} \times g_{Skel_i}\right) + \varepsilon}{\sum_i \left(p_{Skel_i}\right) + \sum_i \left(g_{Skel_i}\right) + \varepsilon} + \theta \left| \frac{Len(g_{Skel}) - Len(p_{Skel})}{Len(g_{Skel})} \right| \tag{4}$$

where g_{Skel} represents the actual road skeleton, p_{Skel} represents the predicted road skeleton, and $Len(\cdot)$ represents the length of the skeleton.

The total loss function is the weighted sum of the above three loss functions. The specific calculation is shown in Equation (5).

$$L_{Total} = k_1 L_{Dice} + k_2 L_{Focal} + k_3 L_{Skeleton} \quad (5)$$

where k_1, k_2, k_3 are adaptive loss weights determined by a weight annealing strategy during training.

k_1 denotes the weight of the Dice Loss, which is used to improve the segmentation overlap and integrity of road areas. k_2 denotes the weight of the Focal Loss, which is used to reduce the impact of class imbalance between road and background. k_3 denotes the weight of the Skeleton Loss, which is used to enhance the continuity and topological accuracy of road centerlines.

In the early stage of training, k_1 and k_2 are set relatively high to prioritize the recall of road regions and alleviate training imbalance. As training proceeds to the later stage, k_3 is gradually increased to focus on optimizing the road centerline structure and improving topological continuity. The annealing schedule of k_1, k_2, k_3 is determined empirically based on the validation set performance to achieve the best balance between region segmentation and structural refinement.

3. Materials and Methods

To ensure a fair and comprehensive comparison, a total of 17 representative methods from different developmental stages are used as baseline models. These baselines include traditional CNN-based architectures, Transformer-oriented frameworks, and recent state-of-the-art specialized networks for road extraction. Brief descriptions of each baseline model are provided below.

FCN-8S (2015) [32]: A pioneering, fully convolutional network for semantic segmentation with skip connections to preserve fine-grained details.

U-Net (2015): A classic symmetric encoder–decoder network widely adopted in medical imaging and remote sensing segmentation tasks.

SegNet (2017) [33]: A lightweight encoder–decoder architecture that utilizes max-pooling indices for efficient upsampling.

LinkNet (2017): A lightweight U-Net variant integrated with residual connections to achieve efficient inference.

DeepLabV3 (2017) [34]: A classical semantic segmentation model using atrous spatial pyramid pooling for multi-scale context aggregation.

D-LinkNet (2018) [35]: An improved LinkNet equipped with dilated convolutions to capture multi-scale road features effectively.

ResU-Net (2018) [36]: A U-Net variant enhanced with residual blocks to support training of deeper networks.

CFPNet (2021) [37]: A road extraction network adopting a cascaded feature pyramid to better detect thin and narrow road targets.

TransUNet (2021) [38]: The first hybrid model that combines Transformer and U-Net to capture long-range global context.

RoadExNet (2023) [39]: A dedicated road extraction network that strengthens contextual feature representation for improved road detection.

RemainNet (2023) [40]: A road extraction model focused on enhancing topological connectivity and structural integrity of road networks.

SwinUNet (2023): A pure Transformer-based U-Net architecture with shifted window self-attention for efficient modeling.

SASwin (2024) [41]: A Swin Transformer-based road extraction method with spatial attention refinement for better road details.

DUSA-UNet (2025) [42]: A dual sparse attention-based U-Net for multi-scale road network feature extraction.

SegRoadv2 (2025) [43]: A hybrid network with deformable self-attention to optimize road connectivity and continuity.

WeavingUnet (2026) [44]: A road extraction model that enhances dense structural representation through horizontal–vertical feature weaving and fusion.

These models are carefully selected to guarantee a comprehensive, up-to-date, and convincing comparison with both classic and the latest advanced approaches.

In addition, this study expanded the U3-Road base model, resulting in two high-performance versions: U3-Road-x2 and U3-Road-x4. Both versions deepen the network through serial stacking, doubling or quadrupling the encoder modules and the bridge module, respectively. For example, in U3-Road-x2, an additional set of Res_block_1, Res_block_2, and Swin-T modules is connected in series at each layer of the decoder, and another layer of the same structure is serially stacked in the bridge part. Since the input and output dimensions of all modules are the same, serial stacking can be performed directly for convenient expansion. By moderately deepening the network, the model can better learn the feature differences between the road and the background, further improving segmentation accuracy. Moderate deepening yields more accurate prediction results when sufficient computing power is available, but it also significantly increases the number of model parameters and computational cost. Therefore, blindly deepening the network is not the optimal strategy. In actual engineering deployment, it is necessary to comprehensively consider the accuracy gains and computational costs based on hardware resources and scene requirements.

3.1. Experimental Platform

The experimental platform of this study uses the Ubuntu 24.04 LTS operating system. The processor is an Intel Xeon Platinum 8480+, the memory is 512 GB DDR5-4800 ECC REG, and the GPU is an NVIDIA H20 with single-precision computing power of 44 TFLOPs and half-precision computing power of 148 TFLOPs.

3.2. Dataset

The DeepGlobe road dataset [45] was released by the 2018 DeepGlobe Road Extraction Challenge and consists of high-resolution remote sensing images with a spatial resolution of 0.5 m per pixel. It contains 6226 groups of 1024 × 1024 pixel images. In this study, 70% of the images are allocated to the training set, 15% to the validation set, and 15% to the test set. Examples from the dataset are shown in Figure 6. The DeepGlobe Road dataset is publicly available at http://deepglobe.org/, accessed on 11 May 2026.

The dataset from Massachusetts [46] is a high-quality aerial road image dataset with a spatial resolution of 1 m per pixel. This dataset contains 1171 groups of 1500 × 1500 pixel images. To achieve consistent alignment of feature sizes during network downsampling and upsampling, the original images are cropped to 1024 × 1024 pixels, which is a power of 2 and avoids misalignment or unnecessary padding of the feature maps. To reduce spatial redundancy between samples and ensure strict evaluation, only one 1024 × 1024 image block is extracted from each original image. These image blocks are divided into training set, validation set, and test set, with proportions of 70%, 15%, and 15%, respectively. Examples of the cropped dataset are shown in Figure 7. The Massachusetts roads dataset is available at https://www.cs.toronto.edu/~vmnih/data/, accessed on 11 May 2026.

Figure 6. Sample of the DeepGlobedataset: (**a**) satellite image; (**b**) road annotation corresponding to (**a**).

Figure 7. Sample of the Massachusetts dataset: (**a**) satellite image; (**b**) road annotation corresponding to (**a**).

The CHN6-CUG road dataset [47] was created and shared by Wuhan University of Geosciences in China. The spatial resolution is 0.5 m per pixel, and it contains 4511 groups of 512 × 512 pixel images. These images are divided into training, validation, and test sets in proportions of 70%, 15%, and 15%, respectively. Some images of the dataset are shown in Figure 8. The CHN6-CUG dataset is available at http://grzy.cug.edu.cn/zhuqiqi/, accessed on 1 April 2026.

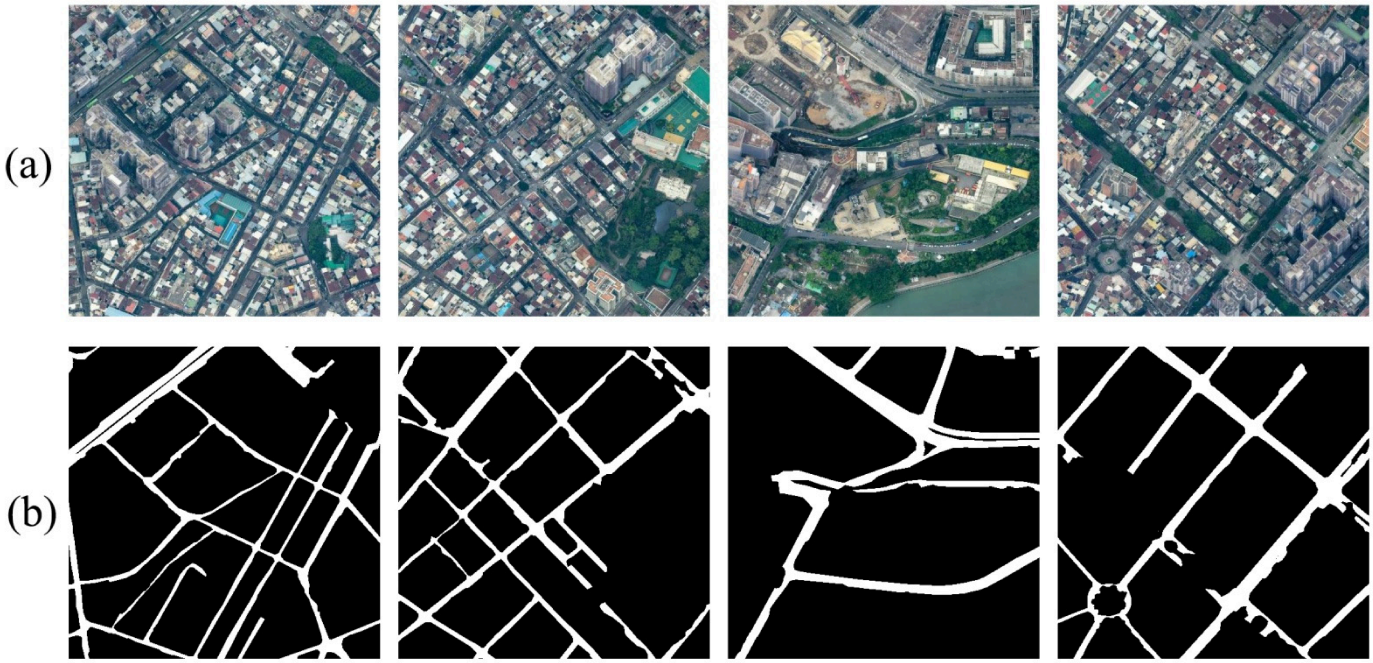

Figure 8. Sample of the CHN6-CUG dataset: (**a**) satellite image; (**b**) corresponding road annotation for (**a**).

3.3. Model Parameter Settings

The U3-Road model employs the Stochastic Gradient Descent (SGD) optimizer [48]. The initial learning rate is set to 0.001, and a polynomial decay strategy is adopted with an exponent of 0.9 and a weight decay coefficient of 0.0001. The batch size is set to 8, and training runs for 100 iterations.

3.4. Evaluation Indicators

To quantitatively evaluate the performance of the U3-Road proposed in this paper, Precision, Recall, F1 score, Accuracy (PA), Intersection over Union (IoU), and Mean Intersection over Union (MIoU) are selected as evaluation indicators for accuracy. The calculation methods are shown in Formulas (6)–(11). Here, TP represents the number of road pixels correctly classified as roads, FP represents the number of background pixels incorrectly classified as roads, FN represents the number of road pixels incorrectly classified as background, and TN represents the number of background pixels correctly classified as background.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

$$\text{F}_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

$$\text{PA} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{9}$$

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

$$\text{MIoU} = \frac{\text{TP}}{2 \times (\text{TP} + \text{FP} + \text{FN})} + \frac{\text{TN}}{2 \times (\text{TN} + \text{FN} + \text{FP})} \tag{11}$$

4. Experimental Results and Analysis

4.1. Quantitative Result Analysis

The quantitative evaluation results for each model on the three datasets—DeepGlobe, Massachusetts, and CHN6-CUG—are shown in Tables 1–3.

Table 1. Evaluation results of each model on the DeepGlobedataset (%). The best and worst performance values in each row are highlighted in yellow and red, respectively.

Methods	Precision	Recall	F1	PA	IoU	MIoU
FCN-8S	70.52	77.70	73.94	94.74	58.65	76.48
SegNet	70.80	75.17	72.92	94.64	57.38	75.80
U-Net	70.06	78.84	74.19	94.73	58.97	76.64
LinkNet	76.13	74.41	75.26	95.30	60.34	77.64
RoadExNet	75.42	77.23	76.31	95.40	61.70	78.36
RemainNet	77.43	77.19	77.31	95.65	63.02	79.16
D-LinkNet	74.55	78.60	76.52	95.37	61.97	78.48
CFPNet	80.97	76.83	78.85	96.04	65.08	80.40
DeepLabV3	78.39	82.73	80.50	96.15	67.36	81.59
ResUnet	81.31	80.32	80.81	96.34	67.80	81.92
TransUNet	82.89	80.24	81.54	96.51	68.84	82.53
SwinUNet	81.56	81.28	81.42	96.44	68.66	82.40
SASwin	83.74	79.78	81.71	96.57	69.08	82.68
DUSA-UNet	82.65	79.18	80.88	96.41	67.90	82.00
SegRoadv2-l	84.25	77.54	80.75	96.45	67.72	81.94
WeavingUnet	82.99	79.28	81.09	96.45	68.20	82.18
U3-Road-base	82.86	82.50	82.68	96.68	70.47	83.43
U3-Road-x2	84.20	83.53	83.86	96.91	72.21	84.43
U3-Road-x4	87.26	83.14	85.15	97.22	74.15	85.56

Table 2. Evaluation results of each model on the Massachusetts dataset (%). The best and worst performance values in each row are highlighted in yellow and red, respectively.

Methods	Precision	Recall	F1	PA	IoU	MIOU
FCN-8S	69.43	78.77	73.81	94.90	58.49	76.50
SegNet	68.73	76.03	72.20	94.66	56.49	75.38
U-Net	69.45	77.72	73.35	94.85	57.92	76.19
LinkNet	75.14	78.78	76.92	95.69	62.49	78.93
RoadExNet	72.78	76.50	74.60	95.25	59.48	77.19
RemainNet	73.31	78.03	75.60	95.41	60.77	77.91
D-LinkNet	79.21	77.24	78.21	96.08	64.22	80.00
CFPNet	74.56	78.99	76.71	95.63	62.22	78.75
DeepLabV3	81.13	76.67	78.84	96.25	65.07	80.52
ResUnet	81.20	77.49	79.30	96.31	65.70	80.87
TransUNet	79.92	77.25	78.56	96.16	64.70	80.28

Table 2. *Cont.*

Methods	Precision	Recall	F1	PA	IoU	MIOU
SwinUNet	81.12	78.31	79.69	96.36	66.24	81.16
SASwin	81.35	79.90	80.62	96.50	67.53	81.88
DUSA-UNet	82.09	77.90	79.94	96.44	66.58	81.37
SegRoadv2-l	83.02	77.79	80.32	96.52	67.11	81.69
WeavingUnet	80.38	81.29	80.83	96.48	67.83	82.02
U3-Road-base	81.66	82.43	82.04	96.71	69.55	83.00
U3-Road-x2	83.83	82.57	83.20	96.96	71.23	83.97
U3-Road-x4	84.37	84.43	84.40	97.15	73.01	84.97

Table 3. Evaluation results of each model on the CHN6-CUG dataset (%). The best and worst performance values in each row are highlighted in yellow and red, respectively.

Methods	Precision	Recall	F1	PA	IoU	MIoU
FCN-8S	76.98	74.22	75.58	91.82	60.74	75.69
SegNet	76.43	75.17	75.79	91.81	61.02	75.81
U-Net	78.85	76.21	77.51	92.46	63.28	77.30
LinkNet	76.86	77.00	76.93	92.13	62.51	76.72
RoadExNet	79.98	76.99	78.46	92.79	64.55	78.13
RemainNet	80.61	75.58	78.02	92.74	63.96	77.81
D-LinkNet	80.65	77.69	79.14	93.02	65.48	78.72
CFPNet	83.19	77.13	80.05	93.44	66.73	79.59
DeepLabV3	83.66	75.47	79.35	93.30	65.77	79.04
ResUnet	83.77	77.29	80.40	93.57	67.22	79.91
TransUNet	82.48	78.12	80.24	93.44	67.00	79.71
SwinUNet	87.05	76.31	81.33	94.03	68.53	80.83
SASwin	87.51	76.52	81.64	94.13	68.98	81.12
DUSA-UNet	85.47	76.63	80.81	93.79	67.80	80.33
SegRoadv2-l	87.75	74.53	80.60	93.88	67.50	80.25
WeavingUnet	85.74	77.15	81.22	93.92	68.38	80.69
U3-Road-base	89.32	77.44	82.96	94.57	70.87	82.31
U3-Road-x2	90.80	77.15	83.42	94.77	71.56	82.77
U3-Road-x4	91.84	79.37	85.15	95.28	74.14	84.34

As shown in Table 1, for the DeepGlobe dataset, most evaluation metrics for the U3-Road model surpass those of all comparison models. Specifically, the IoU of U3-Road-base reaches 70.47%, which is 1.39% higher than SASwin, 11.77% higher than the classic U-Net, and 10.13% higher than the LinkNet method. This clearly demonstrates the superiority of the U3-Road model in extracting mountainous rural roads. In terms of specific indicators, the Precision and Recall of the proposed U3-Road model reach 82.86% and 82.50%, respectively, with an F1 score of 82.68%, indicating that the model achieves good balance between reducing false positives and false negatives. The performance of U3-Road-x2 and U3-Road-x4 shows that deepening and stacking the model can further improve

performance and is suitable for fine-grained road extraction tasks, but this requires more hardware resources.

For the Massachusetts dataset, the U3-Road model maintains a leading advantage. The IoU of the U3-Road-base model reaches 69.55%, which is 1.72% higher than WeavingUnetand 11.63% higher than U-Net. The performance of U3-Road-x2 and U3-Road-x4 is also better than that of U3-Road-base. This dataset mainly consists of urban green space scenes, with dense road networks, many intersections, and building obstructions. The modules of the U3-Road model effectively enhance multi-scale feature fusion, improving recognition of dense roads and obscured areas. According to the Recall metric, the U3-Road model achieves 82.43%, which is 6.4% higher than SegNet, indicating that the model can better capture narrow roads and partially obscured road segments. The Precision metric reaches 81.66%, ensuring the reliability of the prediction results and reducing false detections of non-road areas such as building shadows and vegetation.

The CHN6-CUG dataset represents a Chinese urban scene, featuring complex terrain and traffic conditions. The U3-Road-base model achieved an IoU of 70.87% on this dataset, which was 1.91% higher than that of SASwin and 7.59% higher than that of U-Net. This verifies the applicability of the U3-Road model for extracting extensive urban roads. Some areas inthis dataset have significant variations in road width and blurred road and edge boundaries. The U3-Road model effectively enhanced its adaptability to high-quality images and complex road shapes by introducing an attention mechanism and a multi-scale feature fusion module. From the Precision metric, the U3-Road-base model reached 89.32%, which was 1.81% higher than that of the best-performing model in the comparison, indicating that the model has higher accuracy in extracting road boundaries in complex terrains and can better adapt to a road network layout with Chinese characteristics.

By combining the quantitative results from the three datasets, it can be concluded that the U3-Road model performs optimally in most evaluation metrics, with an average IoU of 70.30%, which is 1.77% higher than that of SASwin. This demonstrates the model's strong generalization ability and extraction accuracy across different resolutions, geographical environments, and road types. Classical deep learning methods, such as U-Net and FCN-8S, although capable of basic road extraction, have limited accuracy in complex scenarios. SwinUNet and SASwin, by introducing Transformer attention mechanisms or high-resolution feature preservation strategies, have significantly improved performance. The U3-Road model, by further optimizing feature fusion and boundary refinement, has achieved even higher extraction accuracy.

4.2. Qualitative Result Analysis

To present the road extraction results of each model more intuitively, three typical scene samples from different datasets were selected for qualitative comparison. The results are shown in Figures 9–17 (the red rectangles in the figures are used to highlight the inconsistent regions between the predicted segmentation results and the ground-truth labels, where the extracted road areas are incomplete, missing, or mismatched the manual annotations), corresponding to the mountain–rural scene from the DeepGlobe dataset, the urban dense road scene from the Massachusetts dataset, and the urban road scene from the CHN6-CUG dataset. Each scene includes the original image, the true label, the prediction results of the U3-Road model, and several comparison models.

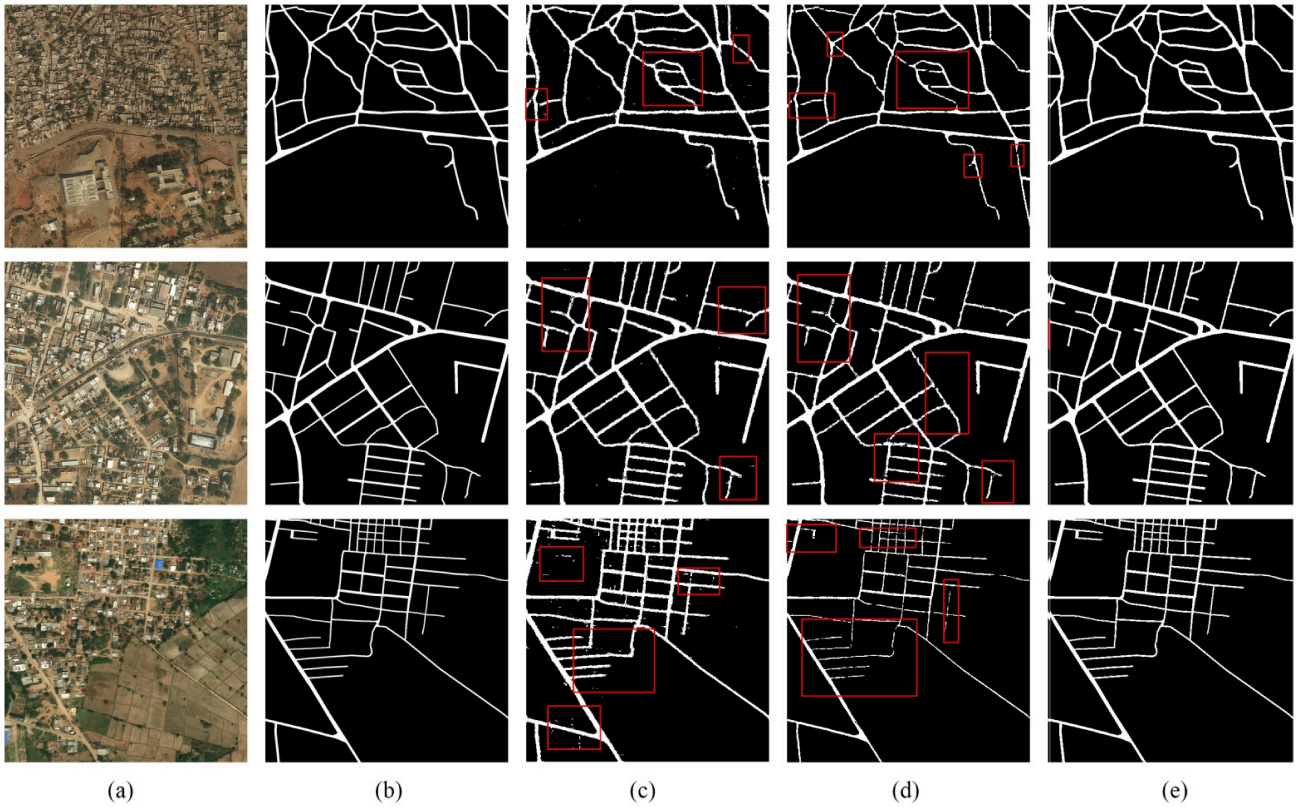

Figure 9. The prediction results on the DeepGlobe dataset: (**a**) original image; (**b**) true label; (**c**) U-Net; (**d**) LinkNet; (**e**) U3-Road-base.

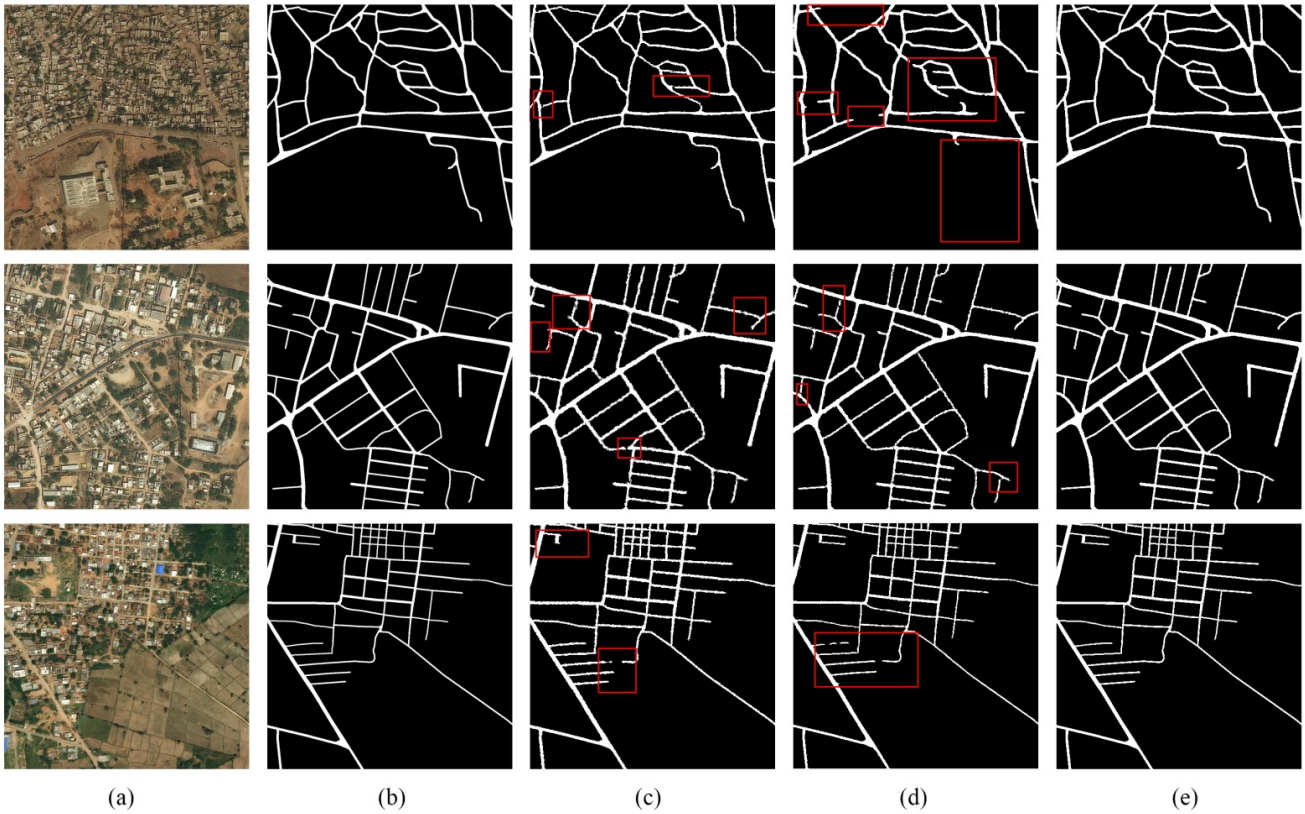

Figure 10. The prediction results on the DeepGlobe dataset: (**a**) original image; (**b**) true label; (**c**) SwinUNet; (**d**) SASwin; (**e**) U3-Road-base.

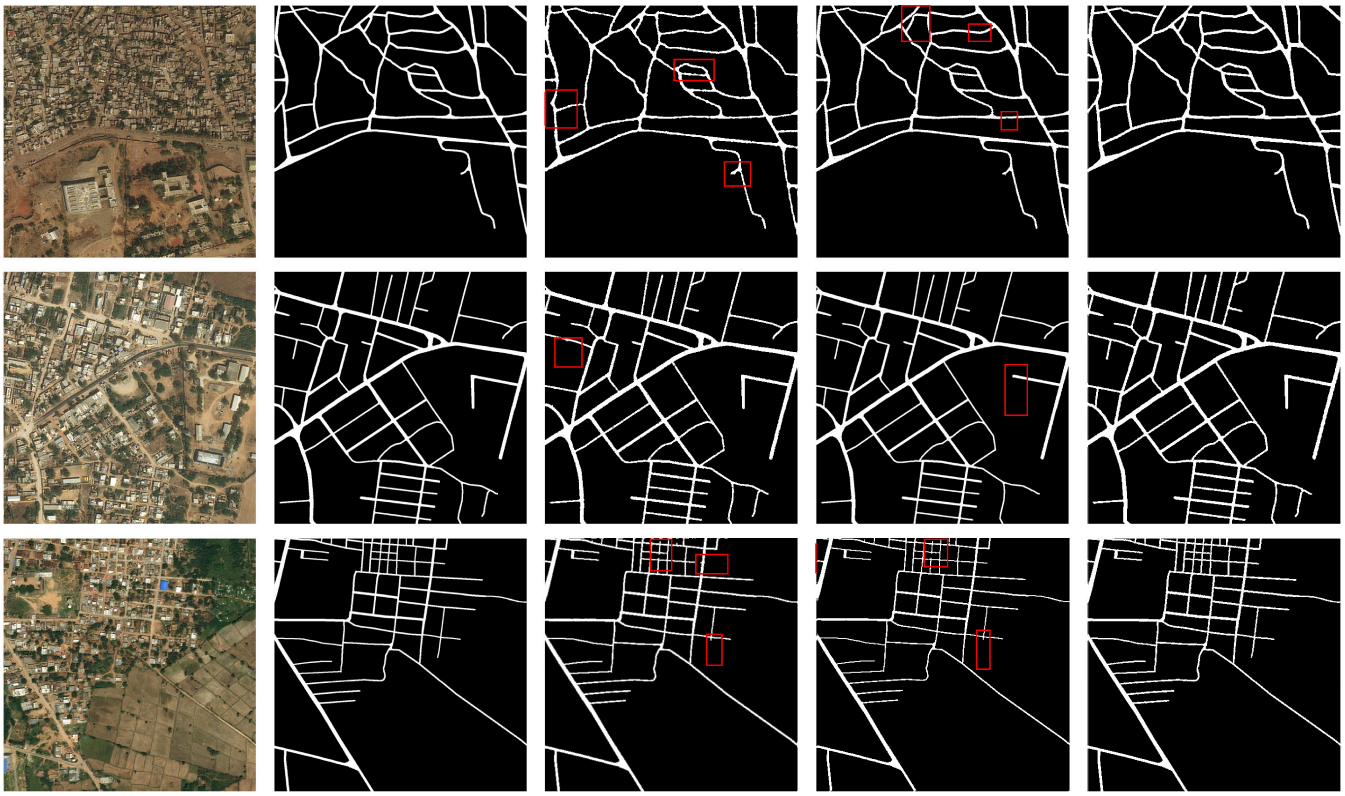

Figure 11. The prediction results on the DeepGlobe dataset: (**a**) original image; (**b**) true label; (**c**) DUSA-UNet; (**d**) WeavingUnet; (**e**) U3-Road-base.

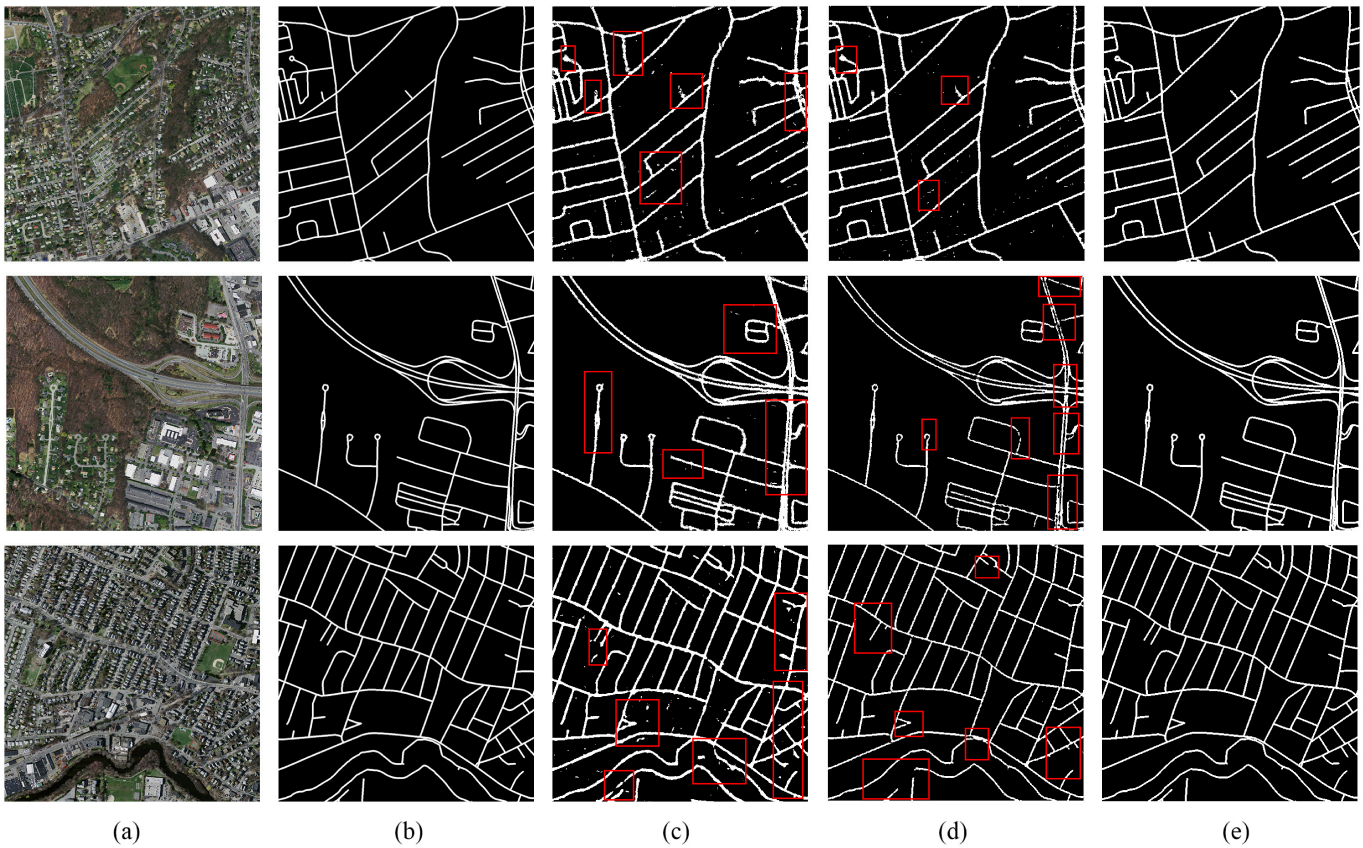

Figure 12. The prediction results on the Massachusetts dataset: (**a**) original image; (**b**) true label; (**c**) U-Net; (**d**) LinkNet; (**e**) U3-Road-base.

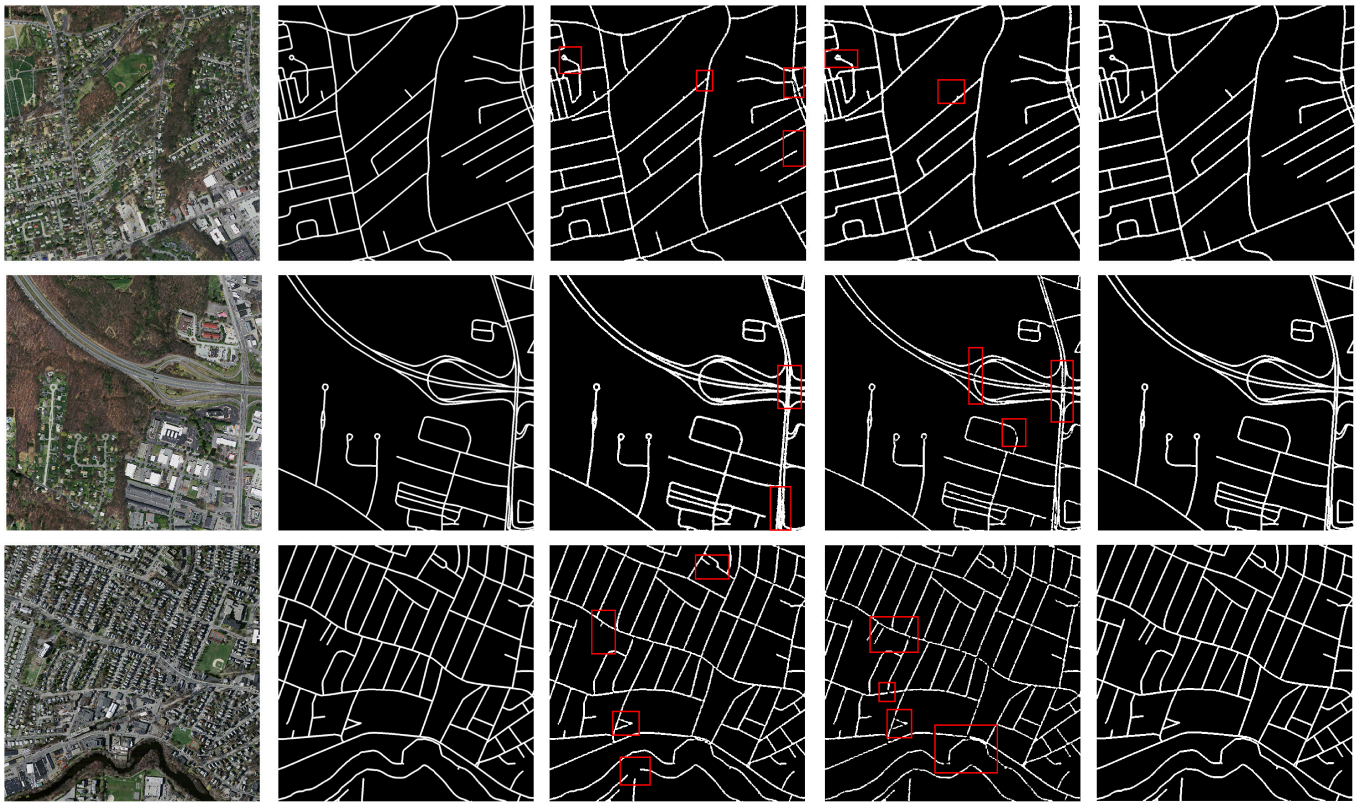

Figure 13. The prediction results on the Massachusetts dataset: (**a**) original image; (**b**) true label; (**c**) SwinUNet; (**d**) SASwin; (**e**) U3-Road-base.

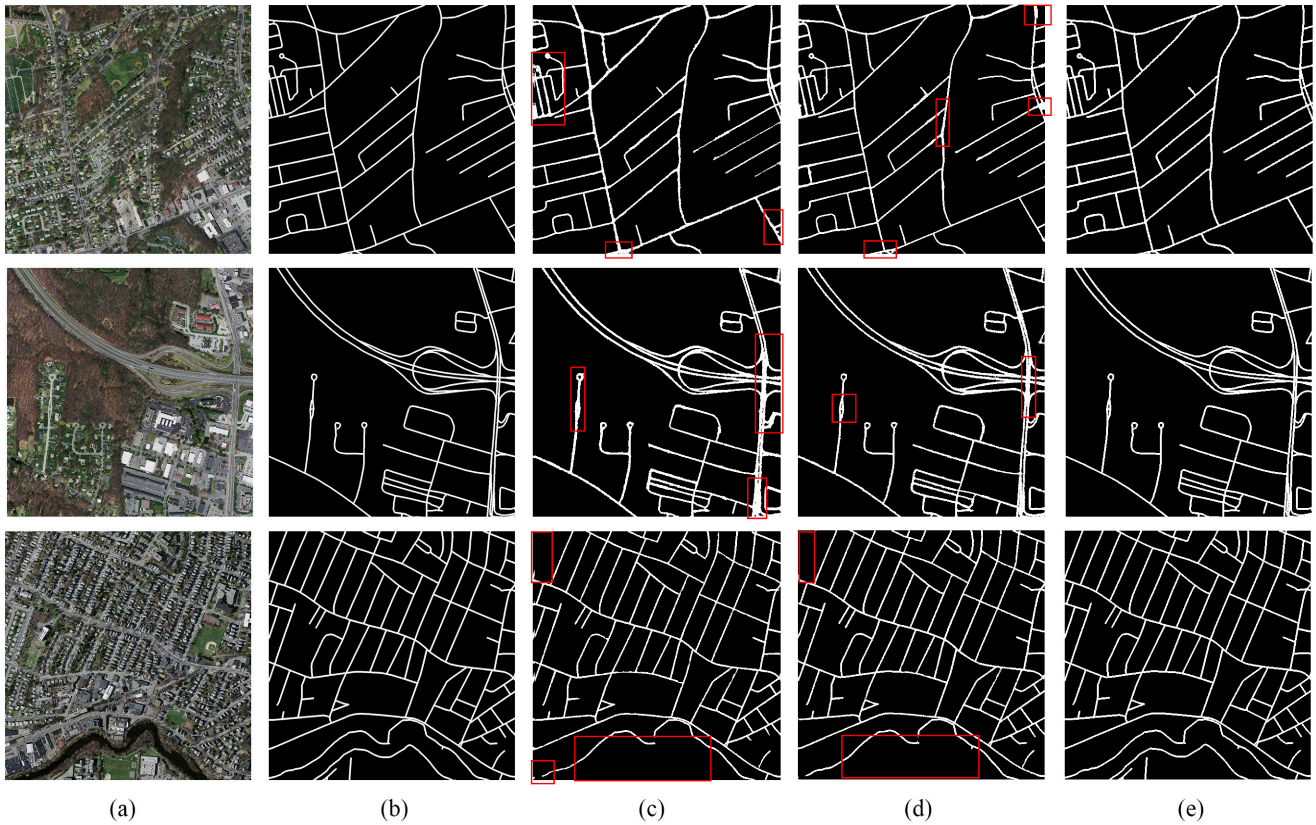

Figure 14. The prediction results on the Massachusetts dataset: (**a**) original image; (**b**) true label; (**c**) DUSA-UNet; (**d**) WeavingUnet; (**e**) U3-Road-base.

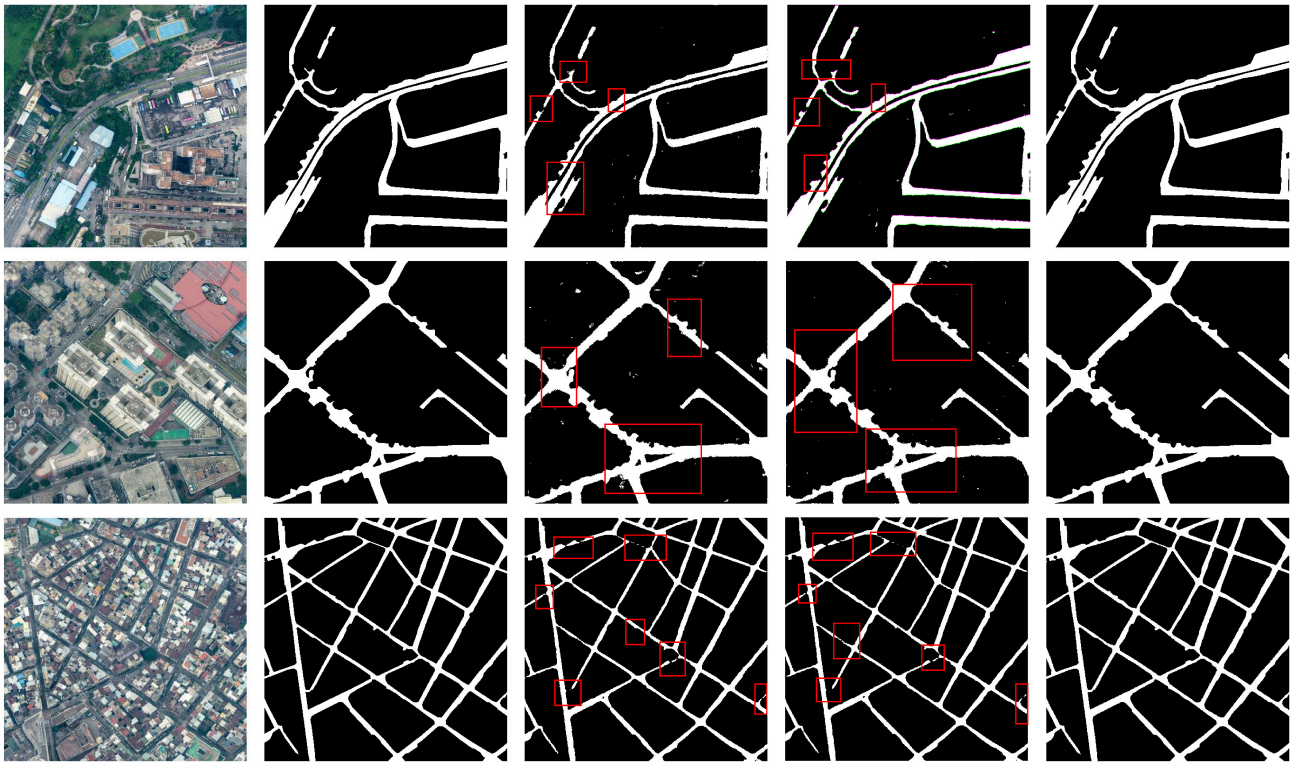

Figure 15. The prediction results on the CHN6-CUG dataset: (**a**) original image; (**b**) true label; (**c**) U-Net; (**d**) LinkNet; (**e**) U3-Road-base.

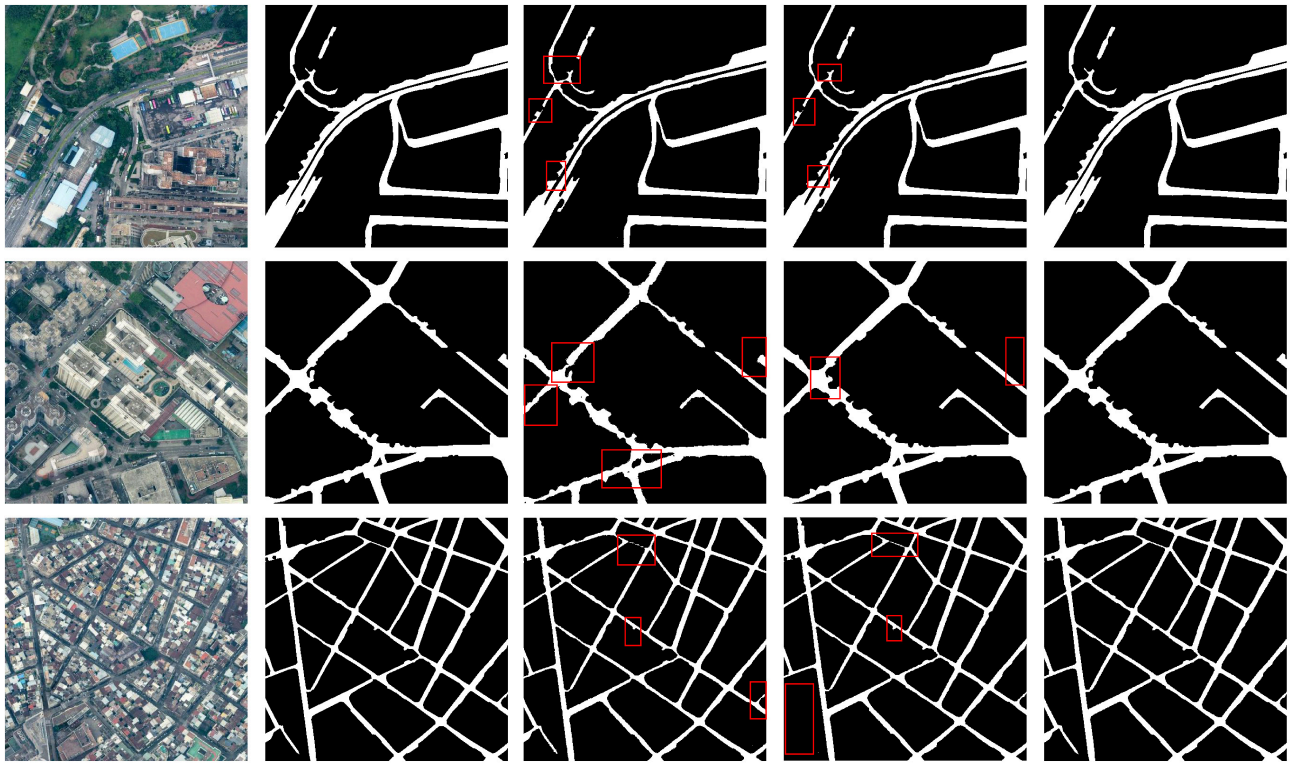

Figure 16. The prediction results on the CHN6-CUG dataset: (**a**) original image; (**b**) true label; (**c**) SwinUNet; (**d**) SASwin; (**e**) U3-Road-base.

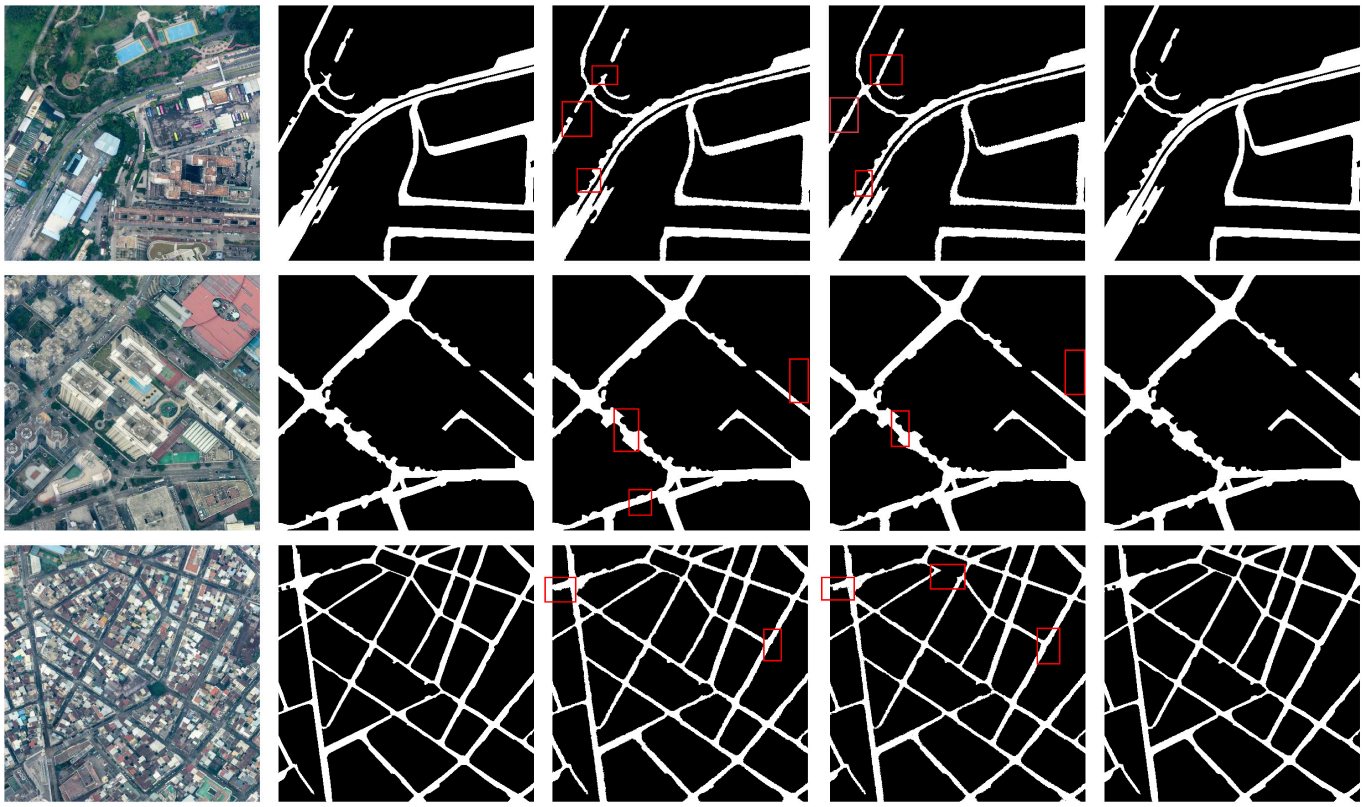

Figure 17. The prediction results on the CHN6-CUG dataset: (**a**) original image; (**b**) true label; (**c**) DUSA-UNet; (**d**) WeavingUnet; (**e**) U3-Road-base.

As shown in Figure 9, in the mountainous rural scenes of the DeepGlobe dataset, the original images present challenges such as narrow rural roads, vegetation obstruction, and terrain undulations. The U-Net and LinkNet models can extract the main roads, but their ability to recognize narrow roads and road segments obscured by vegetation is limited, and the boundaries show noticeable jagged edges. The U3-Road model can fully extract all roads, including narrow rural roads and partially obscured segments. The road boundaries are smooth and continuous, closely matching the ground-truth labels, with no obvious false positive or false negative predictions.

As shown in Figure 10, SwinUNet and SASwin achieve significantly better prediction performance than U-Net and LinkNet, yet still exhibit noticeable road discontinuities and missed detections.

As shown in Figure 11, DUSA-UNet and WeavingUnet can predict road details with relative accuracy. However, burr-like artifacts still appear in the prediction results for road segments occluded by ground objects, and some narrow roads, especially road ends, fail to be detected successfully.

As shown in Figure 12, roads in the Massachusetts dataset are relatively narrow and densely distributed. Although most roads are clearly distinguishable from surrounding ground objects, some road segments remain particularly challenging to identify. U-Net and LinkNet can extract most roads, but they tend to over-detect by misclassifying certain non-road regions as road pixels. Furthermore, these models perform unsatisfactorily in handling fine-grained details. Specifically, they fail to separate adjacent roads individually when multiple roads are close together, often resulting in severe road adhesion.

As shown in Figure 13, SwinUNet and SASwin achieve more accurate road extraction results than U-Net and LinkNet on the Massachusetts dataset. Benefiting from their strong

global context modeling capabilities, these two models effectively alleviate over-detection and reduce false predictions in non-road regions. However, they still experience slight missed detections, especially for narrow and small road segments. Although road adhesion is partially mitigated, incomplete extraction of fine-grained road structures remains evident in the prediction results.

As shown in Figure 14, DUSA-UNet and WeavingUnet perform inadequately in processing fine-grained details. Many challenging road segments remain largely undetected by these two models. In contrast, the proposed U3-Road-base achieves significantly better road extraction performance than the aforementioned methods, yielding more complete and accurate segmentation results.

As shown in Figure 15, roads in the CHN6-CUG dataset are relatively wide, and their two sides are often asymmetric with abundant, diverse details. UNet and LinkNet generally identify the main road regions but fail to correctly detect road segments occluded by other objects. Additionally, a small number of non-road regions are erroneously classified as roads.

As shown in Figure 16, SwinUNet and SASwin generally perform well on the CHN6-CUG dataset. However, their handling of fine-grained road details is still inadequate. In particular, SASwin misses some isolated road segments.

As shown in Figure 17, DUSA-UNet and WeavingUnet tend to miss some isolated road segments and show insufficient detail refinement on both sides of roads. Furthermore, these models incorrectly fill gaps in road structures. In contrast, the proposed U3-Road-base is nearly free of false positives and missed detections, with only minor inconsistencies at fine road edges compared to the ground truth. This demonstrates the model's adaptability to the complex traffic scenarios in Chinese urban areas.

The qualitative comparison results show that the U3-Road model outperforms the comparison models in road extraction performance across different scenarios. It can effectively handle complex issues such as vegetation occlusion, building shadows, image blurring, changes in road width, and terrain undulation. The extracted roads exhibit high integrity, accurate boundaries, and no obvious false positives or false negatives, fully verifying the model's practical application value.

4.3. Complexity Analysis

To illustrate the efficiency of the U3-Road model structure, a model complexity analysis was performed with 13 other models. The input was consistently set as an RGB image of $1024 \times 1024 \times 3$. The specific parameter count (Param) and floating-point operation count (FLOPs) are shown in Table 4.

Table 4. Comparison of the number of parameters (Param) and the number of floating-point operations (FLOPs) for each model.

Methods	Param (M)	FLOPs (G)
FCN-8S	14.7	103.2
SegNet	14.7	113.8
UNet	15.3	285.6
LinkNet	11.8	95.4
RoadExNet	12.5	112.2
RemainNet	12.8	127.3
D-LinkNet	12.1	105.5

Table 4. *Cont.*

Methods	Param (M)	FLOPs (G)
CFPNet	12.3	224.9
DeepLabV3	25.5	323.7
ResUnet	42.5	355.6
TransUNet	38.2	421.2
SwinUNet	28.2	305.1
SASwin	79.9	511.8
DUSA-UNet	66.9	295.3
SegRoadv2-l	92.2	380.8
WeavingUnet	182.7	705.7
U3-Road-base	6.91	134.2
U3-Road-x2	13.77	267.4
U3-Road-x4	27.49	533.8

As shown in Table 4, comprehensive comparisons of parameter quantity (Param) and floating-point operations (FLOPs) clearly demonstrate the distinct advantages of our proposed U3-Road series. Specifically, the U3-Road-base model requires only 6.91 million parameters and 134.2 billion FLOPs, which are significantly lower than those of almost all comparative methods, including FCN-8S, SegNet, UNet, LinkNet, DeepLabV3, TransUNet, SwinUNet, SASwin, DUSA-UNet, and WeavingUnet. Even its extended versions, U3-Road-x2 and U3-Road-x4, maintain a relatively compact model size and computational cost, while remaining much more lightweight than state-of-the-art networks such as SwinUNet, SASwin, and WeavingUnet.

Notably, in terms of parameter quantity, the U3-Road-base model achieves a significant reduction compared to existing methods, which can be attributed to its moderately sized network construction strategy—instead of blindly increasing network depth and width. A moderately compact network structure not only provides a better balance between parameter quantity and model performance but also effectively mitigates the risk of overfitting. Regarding computational cost, while the FLOPs of U3-Road-base are slightly higher than those of earlier models such as FCN-8S and LinkNet (due to the secondary complexity of its self-attention mechanism and the high-dimensional redundancy of the fully connected network), we have adopted optimization strategies including window attention, dilated convolution, and network dimension compression to reduce computational overhead. As a result, the computational cost of U3-Road-base is significantly lower than that of attention mechanism-integrated models such as TransUNet and SASwin, while achieving superior performance.

Collectively, these results confirm that our U3-Road series enables highly efficient road extraction with an ultra-lightweight architecture and low computational complexity, making it more suitable for practical deployment in real-time and resource-constrained scenarios.

4.4. Ablation Experiment

To further verify the effectiveness of each module in the U3-Road model proposed in this paper, an ablation experiment was conducted on the DeepGlobe dataset, and the impact of each module on model performance was evaluated. The specific results are shown in Table 5.

Table 5. Ablation study for the U3-Road model: A is the standard convolution, B is the dilated convolution, C is the Swin Transformer, D is the Transformer Encoder, and E is the composite loss function. The symbol '√' denotes that the corresponding module is adopted in the network.

A	B	C	D	E	U3-Road-Base	
					IoU	MIoU
√					60.75	77.71
√	√				63.89	79.74
√	√	√			67.73	81.90
√	√	√	√		69.34	82.84
√	√	√	√	√	70.47	83.43

Table 5 shows that each functional module of the U3-Road model significantly contributes to improve the IoU metric. With the standard convolution architecture, the model achieves an IoU of 60.75%, which is comparable to the performance of LinkNet. Introducing dilated convolution and adopting the dual-path inference mechanism increases the IoU by 3.14%. Further integration of the Swin Transformer module to construct three-path inference architecture results in an additional 3.84% increase in IoU. Embedding the Transformer encoder into the bridge module yields a further 1.61% improvement. Finally, replacing the single Dice Loss with a composite loss function consisting of Dice Loss, Focal Loss, and Skeleton Loss leads to an additional 1.13% increase in IoU. These results fully validate the rationality and effectiveness of each module's design in the U3-Road model.

To avoid redundancy, the ablation studies in this paper focus on the proposed model rather than all compared methods. Further analysis of other networks can be conducted in future work.

5. Discussion

Extracting high-quality road networks from remote sensing images remains a challenging task due to object occlusion, background interference, spectral variations, and complex road topologies. Traditional lightweight CNN-based models such as U-Net and LinkNet excel at capturing local road details, including edges, textures, and narrow segments. However, they lack effective global context modeling, often leading to fragmented roads, discontinuous centerlines, and false detections in complex scenes. Although pure Transformer-based models can capture long-range dependencies, they usually require heavy computational resources and struggle to preserve fine-grained local details efficiently.

Hybrid CNN-Transformer architectures have emerged as a promising solution by combining the local feature extraction ability of CNNs and the global modeling capability of Transformers. Such frameworks satisfy the dual demands of remote sensing road extraction: precise pixel-level localization and reliable topological integrity. Nevertheless, many existing hybrid models, including SwinUNet, SASwin, DUSA-UNet, SegRoadv2-l, and WeavingUnet, either deepen network layers excessively, rely on heavy window-based or shifted-window attention mechanisms, or focus heavily on topological connectivity optimization. As a result, these models often achieve improved structural continuity at the cost of increased parameters, higher computational complexity, or reduced boundary precision in general pixel-level segmentation.

The experimental results indicate that some recently proposed road extraction models have not comprehensively outperformed classic models (e.g., TransUNet, SwinUNet, and SASwin) across all indicators. This is because classic segmentation models already possess mature and stable feature expression capabilities in remote sensing road extraction tasks,

whereas some new models focus more on structural optimization, such as road topological connectivity and centerline continuity. As a result, they may not necessarily have comprehensive advantages in general pixel-level segmentation indicators.

This study focuses on the accuracy of pixel-level segmentation, as more precise pixel-level prediction results can provide a more reliable foundation for subsequent road connectivity completion and topological repair. Conversely, if a model prioritizes road connectivity in the initial stage, it will often sacrifice local details and boundary accuracy. Subsequently, when supplementing details based on the already connected but information-deficient road structure, the difficulty increases significantly, and errors are more likely to be introduced.

Therefore, a comprehensive comparison between new and existing models can more objectively reflect the effectiveness and practicality of the method proposed in this paper forpixel-level road extraction tasks.

Although the basic version of U3-Road offers only a minor improvement in performance compared to SASwin, SegRoadv2-l, and WeavingUnet, it has fewer parameters and lower hardware requirements. Even when the module is stacked four times to enhance performance, the number of parameters in our model remains lower than those of these similar models, while its accuracy is significantly higher. In contrast, SASwin, SegRoadv2-l, and WeavingUnet already have large parameter counts. Further stacking would significantly increase their computational cost and memory requirements, limiting their practicality for large-scale remote sensing image processing. Therefore, the proposed model demonstrates better scalability and efficiency, and shows more pronounced advantages in actual high-resolution remote sensing applications.

In this context, this study proposes U3-Road, a lightweight hybrid CNN-Transformer network designed to balance accuracy, efficiency, and scalability. Unlike existing hybrid models that typically use deep encoders, heavy self-attention modules, or fixed serial structures, U3-Road adopts a parallel wide-and-shallow architecture that integrates standard convolution, dilated convolution, window attention, and simplified global attention in a nested and stackable manner. This design reduces redundant computations while preserving multi-scale contextual information. In contrast to SwinUNet and SASwin, which rely on hierarchical window partitioning and complex transformer blocks, U3-Road uses lightweight attention combinations to avoid excessive computational overhead. Compared with DUSA-UNet, SegRoadv2-l, and WeavingUnet, which often optimize specifically for road connectivity or centerline extraction, U3-Road maintains stronger performance in general pixel-level segmentation without sacrificing efficiency.

Furthermore, the consistent input–output dimensions of key modules in U3-Road enable flexible stacking. Experimental results show that two-fold and four-fold stacking improve IoU by 1.37% and 3.47%, respectively, demonstrating favorable scalability without significant efficiency degradation. Achieving such a balance between performance, parameters, and FLOPs is difficult for many existing hybrid models, which typically improve accuracy only by deepening structures or enlarging attention windows.

Despite its advantages, U3-Road still has limitations. It currently lacks cross-image long-range dependency modeling. For large-scale stitched remote sensing images covering hundreds of kilometers, cross-image road continuity cannot be fully exploited. Future work will explore lightweight cross-window or cross-patch attention mechanisms to enable long-range road modeling while maintaining low computational costs.

6. Conclusions

This paper presents U3-Road, a lightweight hybrid CNN-Transformer network for remote sensing image road extraction. Extensive experiments were conducted on the DeepGlobe, Massachusetts, and CHN6-CUG datasets, with comparisons against classic

and state-of-the-art models including U-Net, TransUNet, SwinUNet, SASwin, DUSA-UNet, SegRoadv2-l, and WeavingUnet. Quantitative, qualitative, and ablation analyses validate the effectiveness of the proposed model, leading to the following conclusions:

U3-Road achieves superior performance across all evaluation metrics on three datasets, with a mean IoU of 70.3%, outperforming U-Net by 10.24% and SASwin by 1.77%. This demonstrates its strong generalization ability over diverse road environments, resolutions, and topologies.

Qualitative results confirm that U3-Road effectively mitigates fragmentation, false positives, and boundary blurring caused by vegetation occlusion, shadows, and road width variations. It maintains high road integrity and localization accuracy in urban, rural, and mountainous areas.

Ablation studies verify the effectiveness of the multi-scale feature fusion module, attention enhancement module, and boundary refinement module. Together, these components improve the average IoU by 9.72% compared to the baseline structure.

Unlike many existing hybrid models that achieve better performance by increasing model size and computation, U3-Road maintains competitive accuracy with fewer parameters and lower FLOPs, making it more suitable for real-world and resource-constrained applications. Its unique parallel wide-and-shallow structure, dilated convolution-window attention combination, and stackable module design distinguish it from SwinUNet, SASwin, DUSA-UNet, SegRoadv2-l, and WeavingUnet by better balancing precision, efficiency, and scalability.

In summary, U3-Road provides an effective and efficient solution for remote sensing road extraction by synergizing CNN and Transformer structures while optimizing loss and architecture. It achieves high pixel-level accuracy, topological continuity, and computational efficiency simultaneously, making it suitable for large-scale and practical road extraction tasks.

Author Contributions: Conceptualization, Y.W. and H.Z.; methodology, Y.W. and H.Z.; software, H.Z.; validation, H.Z.; formal analysis, H.Z.; investigation, H.Z.; resources, H.Z.; data curation, H.Z.; writing—original draft preparation, Y.W. and H.Z.; writing—review and editing, Y.W. and H.Z.; visualization, H.Z.; supervision, Y.W.; project administration, Y.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant numbers 42477511 and U1901215, and Natural Science Fund of Guangdong Province, grant number 2021A1515011375.

Data Availability Statement: The code is publicly available at: https://github.com/zhf10/u3-Road.git, accessed on 11 May 2026.

Acknowledgments: The authors sincerely acknowledge the 2018 DeepGlobe Road Extraction Challenge for the DeepGlobe road dataset, the Massachusetts datasetproprietors, and China University of Geosciences in Wuhan for sharing the CHN6-CUG road dataset.

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

References

1. Qiu, M.; Huang, L.; Tang, B.-H. ASFF-YOLOv5: Multielement Detection Method for Road Traffic in UAV Images Based on Multiscale Feature Fusion. *Remote Sens.* **2022**, *14*, 3498. [CrossRef]
2. Bhatti, U.A.; Huang, M.; Li, J.; Bazai, S.U.; Aamir, M. *Deep Learning for Multimedia Processing Applications: Volume Two: Signal Processing and Pattern Recognition*, 1st ed.; CRC Press: Boca Raton, FL, USA, 2023.
3. Liu, R.; Wu, J.; Lu, W.; Miao, Q.; Zhang, H.; Liu, X.; Lu, Z.; Li, L. A Review of Deep Learning-Based Methods for Road Extraction from High-Resolution Remote Sensing Images. *Remote Sens.* **2024**, *16*, 2056. [CrossRef]

4. Zhou, N.; Zhou, M.; Wang, X.; Chen, D.; Shi, W.; Zhu, Y.; Liu, J.; Sui, H. Submeter-Level Global-Scale Road Extraction Based on Limited Labeled Data from Optical Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 4411323. [CrossRef]
5. Liu, P.; Wang, Q.; Yang, G.; Li, L.; Zhang, H. Survey of Road Extraction Methods in Remote Sensing Images Based on Deep Learning. *J. Photogramm. Remote Sens. Geoinf. Sci.* **2022**, *90*, 135–159. [CrossRef]
6. Cai, Y.; Mi, S.; Yan, J.; Peng, H.; Luo, X.; Yang, Q.; Wang, J. An Unsupervised Segmentation Method Based on Dynamic Threshold Neural P Systems for Color Images. *Inf. Sci.* **2022**, *587*, 473–484. [CrossRef]
7. Ghandorh, H.; Boulila, W.; Masood, S.; Koubaa, A.; Ahmed, F.; Ahmad, J. Semantic Segmentation and Edge Detection—Approach to Road Detection in Very High Resolution Satellite Images. *Remote Sens.* **2022**, *14*, 613. [CrossRef]
8. Mortazavi, S.S.; Sharafi, Y.; Teshnehlab, M. A Hybrid Deep Transfer Learning Approach for Road Extraction from High-Resolution Satellite Images Based on D-LinkNet and Fuzzy Systems. *Appl. Soft Comput.* **2026**, *157*, 115448.
9. Soni, Y.; Meena, U.; Mishra, V.K.; Khan, A.A.; Khan, R.A. AM-UNet: Road Network Extraction from High-Resolution Aerial Imagery Using Attention-Based Convolutional Neural Network. *J. Indian Soc. Remote Sens.* **2025**, *53*, 135–147. [CrossRef]
10. Zhang, P.; Yi, W.; Song, Y.; Wu, P.; Hampson, K.; Shemery, A. Gaussian Mixture Segmentation for Managing Deterioration of Large-Scale Road Networks. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 12461–12473. [CrossRef]
11. Shen, Y.; Jing, K.; Sun, K.; Liu, C.; Yang, Y.; Liu, Y. Review of Uneven Road Surface Information Perception Methods for Suspension Preview Control. *Sensors* **2025**, *25*, 5884. [CrossRef] [PubMed]
12. Gu, K.; Wang, Y.; Yang, Z.; Cai, Z.; Chen, G.; Li, M. From Empirical Judgment to Data-Driven Approaches: A Survey of Traffic Reorganization and Management During Urban River-Crossing Corridor Construction. *Appl. Sci.* **2025**, *15*, 10133. [CrossRef]
13. Xiao, T.; Ai, T.; Yu, H.; Yang, M.; Liu, P. A Point Selection Method in Map Generalization Using Graph Convolutional Network Model. *Cartogr. Geogr. Inf. Sci.* **2024**, *51*, 20–40. [CrossRef]
14. Fiorentini, N.; Losa, M. Road Detection, Monitoring, and Maintenance Using Remotely Sensed Data. *Remote Sens.* **2025**, *17*, 917. [CrossRef]
15. Khan, M.J.; Singh, P.P. Advanced Road Extraction Using CNN-Based U-Net Model and Satellite Imagery. *e-Prime-Adv. Electr. Eng. Electron. Energy* **2023**, *5*, 100244. [CrossRef]
16. Sertel, E.; Hucko, C.M.; Kabadayı, M.E. Automatic Road Extraction from Historical Maps Using Transformer-Based SegFormers. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 464. [CrossRef]
17. Chen, Z.; Luo, Y.; Wang, J.; Li, J.; Wang, C.; Li, D. DPENet: Dual-Path Extraction Network Based on CNN and Transformer for Accurate Building and Road Extraction. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *124*, 103510. [CrossRef]
18. Liu, W.; Gao, S.; Zhang, C.; Yang, B. RoadCT: A Hybrid CNN-Transformer Network for Road Extraction from Satellite Imagery. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 2501805. [CrossRef]
19. Tao, J.; Chen, Z.; Sun, Z.; Guo, H.; Leng, B.; Yu, Z.; Wang, Y.; He, Z.; Lei, X.; Yang, J. Seg-Road: A Segmentation Network for Road Extraction Based on Transformer and CNN with Connectivity Structures. *Remote Sens.* **2023**, *15*, 1602. [CrossRef]
20. Soydaner, D. Attention Mechanism in Neural Networks: Where It Comes and Where It Goes. *Neural Comput. Appl.* **2022**, *34*, 13371–13385. [CrossRef]
21. Wang, Z.; Liao, Z.; Zhou, B.; Yu, G.; Luo, W. SwinURNet: Hybrid Transformer-CNN Architecture for Real-Time Unstructured Road Segmentation. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 5035816. [CrossRef]
22. Wang, R.; Cai, M.; Xia, Z.; Zhou, Z. Remote Sensing Image Road Segmentation Method Integrating CNN-Transformer and UNet. *IEEE Access* **2023**, *11*, 144446–144455. [CrossRef]
23. Shafiq, M.; Gu, Z. Deep Residual Learning for Image Recognition: A Survey. *Appl. Sci.* **2022**, *12*, 8972. [CrossRef]
24. Zhang, Q. A Novel ResNet101 Model Based on Dense Dilated Convolution for Image Classification. *SN Appl. Sci.* **2022**, *4*, 9. [CrossRef]
25. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 9992–10002.
26. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems 30 (NeurIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
27. Chaurasia, A.; Culurciello, E. LinkNet: Exploiting Encoder Representations for Efficient Semantic Segmentation. In *Proceedings of the 2017 IEEE Visual Communications and Image Processing (VCIP)*; IEEE: St. Petersburg, FL, USA, 2017; pp. 1–4.
28. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer International Publishing: Cham, Switzerland, 2015; Volume 9351, pp. 234–241.
29. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation. In *Computer Vision—ECCV 2022 Workshops*; Karlinsky, L., Michaeli, T., Nishino, K., Eds.; Springer Nature: Cham, Switzerland, 2023; Volume 13803, pp. 205–218.

30. Kato, S.; Hotta, K. Adaptive T-vMF Dice Loss: An Effective Expansion of Dice Loss for Medical Image Segmentation. *Comput. Biol. Med.* **2024**, *168*, 107695. [CrossRef]
31. Mukhoti, J.; Kulharia, V.; Sanyal, A.; Golodetz, S.; Torr, P.H.S.; Dokania, P.K. Calibrating Deep Neural Networks Using Focal Loss. In Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–12 December 2020; Volume 33.
32. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Boston, MA, USA, 2015; pp. 3431–3440.
33. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
34. Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587. [CrossRef]
35. 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)*; IEEE: Salt Lake City, UT, USA, 2018; pp. 192–1924.
36. Baldeon-Calisto, M.; Lai-Yuen, S. ResU-Net: Residual convolutional neural network for prostate MRI segmentation. In *IISE Annual Conference. Proceedings*; Institute of Industrial and Systems Engineers (IISE): Orlando, FL, USA, 2018; pp. 731–736.
37. Lou, A.; Loew, M. CFPNET: Channel-Wise Feature Pyramid for Real-Time Semantic Segmentation. In *Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP)*; IEEE: Anchorage, AK, USA, 2021; pp. 1894–1898.
38. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv* **2021**, arXiv:2102.04306. [CrossRef]
39. 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]
40. Li, Z.; Chen, H.; Jing, N.; Li, J. RemainNet: Explore Road Extraction from Remote Sensing Image Using Mask Image Modeling. *Remote Sens.* **2023**, *15*, 4215. [CrossRef]
41. Zhu, X.; Huang, X.; Cao, W.; Yang, X.; Zhou, Y.; Wang, S. Road Extraction from Remote Sensing Imagery with Spatial Attention Based on Swin Transformer. *Remote Sens.* **2024**, *16*, 1183. [CrossRef]
42. Song, J.; Sun, Y.; Cai, Z.; Xiao, L.; Huang, Y.; Zheng, Y. DUSA-UNet: Dual Sparse Attentive U-Net for Multiscale Road Network Extraction. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5624514. [CrossRef]
43. Yu, Z.; Chen, Z.; Xiao, K.; Lei, X.; Tang, R.; He, Q.; Guo, H. SegRoadv2: A Hybrid Deformable Self-Attention and Convolutional Network for Road Extraction with Connectivity Structure. *Int. J. Digit. Earth* **2025**, *18*, 2480267. [CrossRef]
44. Ma, X.; Li, J.; Lv, M. WeavingUnet: Enhancing Dense Road Extraction by Integrating Horizontal and Vertical Features. *Int. J. Appl. Earth Obs. Geoinf.* **2026**, *146*, 105037. [CrossRef]
45. Li, Z.; Zou, D.; Tang, J.; Zhang, Z.; Sun, M.; Jin, H. A Comparative Study of Deep Learning-Based Vulnerability Detection System. *IEEE Access* **2019**, *7*, 103184–103197. [CrossRef]
46. Yang, Z.; Zhou, D.; Yang, Y.; Zhang, J.; Chen, Z. Road Extraction from Satellite Imagery by Road Context and Full-Stage Feature. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 8000405. [CrossRef]
47. Jayakumar, T.V.; Mishra, D.; Ramiya, A.M.; Parulekar, B.; Pranav, C.V.; Singla, J.G. Evaluation of UNet Architecture Variants for Road Network Extraction from High Resolution Geospatial Dataset. In *2024 IEEE India Geoscience and Remote Sensing Symposium (InGARSS), Goa, India, 2–4 December 2024*; IEEE: Piscataway, NJ, USA, 2024; pp. 1–4.
48. Dagal, I.; Tanriöven, K.; Nayir, A.; Akın, B. Adaptive Stochastic Gradient Descent (SGD) for Erratic Datasets. *Future Gener. Comput. Syst.* **2025**, *166*, 107682. [CrossRef]

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.