remote sensing

MDPI

Article

RFM-UNet: Hybrid Frequency–Mamba UNet for Remote-Sensing Road Extraction

Pu Song [1], Peng Yu [1,2,*], Xiaojing Zhong [3], Shuizhen Wu [1], Junbin Tong [1], Yuanrong He [1,2], Lujun Zhang [4], Guangchun Li [2] and Mengmeng Li [5]

[1] Big Data Institution of Natural Hazards Monitoring for Digital Fujian, Xiamen University of Technology, Xiamen 361024, China; 2422091014@stu.xmut.edu.cn (P.S.); 2422071045@stu.xmut.edu.cn (S.W.); 2522081008@stu.xmut.edu.cn (J.T.); 2012112001@xmut.edu.cn (Y.H.)
[2] Key Laboratory of Southeast Coast Marine Information Intelligent Perception and Application, Ministry of Natural Resources, Zhangzhou 363005, China; guangchunlee@163.com
[3] Xiamen Key Laboratory of Green and Smart Coastal Engineering, College of Harbour and Coastal Engineering, Jimei University, Xiamen 361021, China; xjzhong@jmu.edu.cn
[4] Shanghai Investigation, Design and Research Institute, Shanghai 200125, China; zhang_lujun@ctg.com.cn
[5] Key Laboratory of Spatial Data Mining and Information Sharing of Ministry of Education, The Academy of Digital China, Fuzhou University, Fuzhou 350108, China; mli@fzu.edu.cn
* Correspondence: yupeng@xmut.edu.cn

Highlights

What are the main findings?

- We propose RFM-UNet, a hybrid Frequency–Mamba network featuring an ADAMamba encoder that jointly captures local geometries and global dependencies at linear computational complexity.
- By integrating a Multi-Scale Adaptive Fusion Module (MAFM) and a Dual-Spectrum Aggregation Module (DualSpec), the network effectively preserves narrow-road connectivity and mitigates shadow occlusions, achieving a superior performance–efficiency balance across three public datasets.

What are the implications of the main findings?

- The success of the ADAMamba block demonstrates that rigorous topological continuity in remote-sensing imagery can be achieved efficiently without the quadratic computational bottleneck of traditional Vision Transformers.
- The proposed spatial–frequency synergy offers a highly robust approach for extracting continuous structures under severe occlusions, significantly reducing the need for manual post-processing in applications like GIS updating and autonomous navigation.

Abstract

Road-network extraction from very high-resolution (VHR) remote-sensing imagery remains a challenging task owing to the structural sparsity, topological complexity, and severe occlusions of road networks. Conventional graph-based approaches preserve topological consistency yet incur considerable computational overhead, whereas prevailing convolutional neural network (CNN) and Transformer architectures struggle to reconcile long-range contextual modeling with computational efficiency. To address these limitations, this study proposes RFM-UNet, a hybrid frequency and state–space network designed for road-network segmentation. Specifically, the encoder integrates Mamba blocks with an Anisotropic Directional Attention (ADA) module to jointly capture local geometric cues and global dependencies at linear computational complexity. In addition, a Multi-Scale Adaptive Fusion Module (MAFM) is introduced to dynamically recalibrate multi-stage features, thereby suppressing cross-scale interference and preserving the connectivity of narrow

Check for updates

Academic Editor: Taskin Kavzoglu

Received: 4 June 2026
Revised: 10 July 2026
Accepted: 22 July 2026
Published: 3 August 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.

roads. To enhance robustness against shadow-induced occlusions, a Dual-Spectrum Aggregation Module (DualSpec) decouples the phase and amplitude spectra in the frequency domain and fuses them with spatial features, effectively mitigating spurious responses and background noise characterized by similar textures. Quantitative and qualitative experiments on three public datasets demonstrate that RFM-UNet consistently outperforms current state-of-the-art methods.

Keywords: road extraction; remote sensing; Mamba; frequency-domain; feature fusion

1. Introduction

Very high-resolution (VHR) remote-sensing imagery serves as a foundational data source for geospatial extraction, particularly for mapping critical infrastructure such as road networks. Accurate road extraction is essential for urban planning, traffic management, and disaster emergency response [1]. Traditionally, map updating heavily relied on manual interpretation [2], a process that is inherently labor-intensive and lacks scalability. Although conventional computer vision methods have been introduced to automate this process, they frequently struggle to handle variations in road appearance across different scenes and scales. Consequently, automating road extraction from VHR imagery remains a complex and challenging research endeavor [3]. In recent years, the rapid development of deep-learning (DL) techniques has provided a promising avenue for addressing this issue. By learning hierarchical features and contextual information from imagery, DL models are better equipped to handle the diversity of road characteristics, thereby advancing automated road extraction significantly.

Within the deep-learning paradigm, Convolutional Neural Networks (CNNs) have become the predominant architecture for road extraction due to their robust hierarchical feature extraction capabilities [4,5]. However, the intrinsic locality of convolution operations restricts the receptive field, thereby constraining the modeling of long-range spatial dependencies. This locality limitation frequently yields fragmented topologies in complex scenes characterized by occlusions, shadows, and contextual interference [6,7]. To overcome these limitations, Vision Transformers (ViTs) [8] and hybrid architectures utilize self-attention mechanisms to capture long-range dependencies, outperforming conventional CNNs in global feature modeling [9–11]. Nevertheless, as illustrated in Figure 1a,b, the quadratic computational complexity of self-attention and its tendency to overlook fine-grained spatial details impede the precise delineation of narrow, continuous road networks. Therefore, reconciling global context modeling with precise local feature representation remains a primary bottleneck in VHR road extraction [12,13].

Recently, State–Space Models (SSMs), particularly the Mamba architecture [14], have garnered significant attention for modeling global dependencies with linear computational complexity. Their adaptations, such as Vision Mamba [15] and VMamba [16], have demonstrated substantial potential in computer vision tasks. Leveraging these attributes, Mamba-based architectures are highly suitable for VHR image analysis in remote sensing, as illustrated in Figure 1c. Several recent RS-specific variants, including RS-Mamba [17], RS^{3}Mamba [18], and CM-UNet [19] have validated this potential. However, similar to ViTs, Mamba's sequential processing nature introduces inherent limitations when applied to visual data [20]. Specifically, the unordered nature of spatial relationships among pixels and the prevalence of localized spatial patterns challenge this strictly sequential paradigm, thereby constraining effective spatial feature integration and parallel processing efficiency.

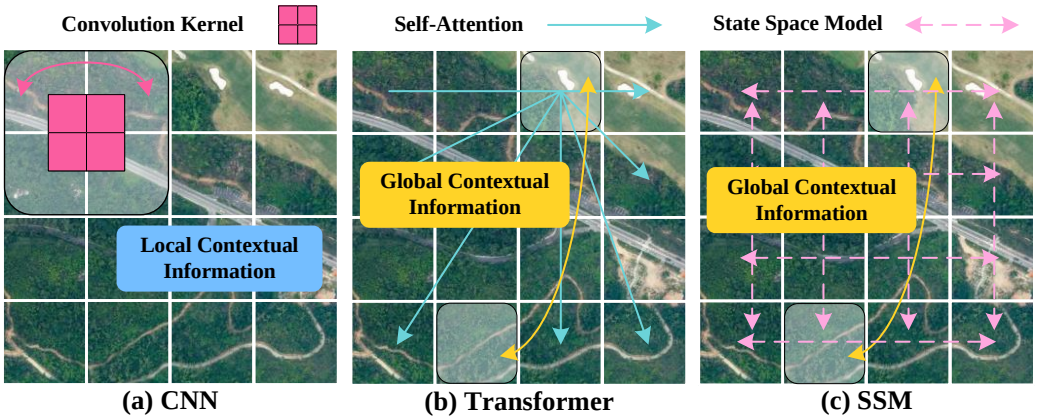


Figure 1. Illustration of different architectures for VHR images.

Beyond architectural advancements, the effective fusion of multi-level features remains a fundamental bottleneck in VHR road extraction [21,22]. Conventional strategies (e.g., simple concatenation, element-wise addition, or bilinear pooling) inadequately bridge the representational gap between deep semantic features and shallow, fine-grained details, thereby limiting their capacity to delineate narrow, elongated, and complex road structures. This necessitates dedicated bridging mechanisms to align high-level semantics with low-level spatial details prior to fusion. Recent efforts have sought to address this gap [23]; for instance, Xu et al. [24] introduced a coarse-to-fine interaction strategy, while flow-based alignment and deformable convolutions [25] have been employed for spatial pre-alignment. However, these intricate multi-scale interactions typically incur substantial computational overhead, thereby hindering practical efficiency.

Moreover, complex backgrounds and uneven illumination frequently induce feature blurring and spatial aliasing in the spatial domain [26], impeding precise road delineation [27]. The Fourier transform decomposes an image into phase and amplitude spectra; the former encapsulates structural semantics, whereas the latter encodes texture and background styles [28]. This distinction highlights their complementary roles in frequency-domain road modeling (Figure 2). Frequency-domain transformation techniques, such as the Discrete Wavelet Transform (DWT) [29] and the Fast Fourier Transform (FFT) [23], exploit this property to enhance structural details while suppressing background noise. Nevertheless, most existing approaches process spatial and frequency information in isolation, overlooking their potential synergy. This limitation significantly restricts their capability to resolve weak boundaries and suppress spurious responses in complex road scenes [30].

To address these limitations, this study proposes RFM-UNet, an efficient hybrid frequency–state–space network tailored for road extraction from VHR imagery. The framework synergistically couples global context modeling with local structural representation under linear computational complexity. The main contributions of this work are summarized as follows:

1. We propose an ADAMamba encoder to reconcile local structural details with global context modeling without incurring quadratic computational complexity. By integrating an Anisotropic Directional Attention (ADA) module with 2D cross-directional scanning, it jointly captures local geometric anisotropy and global semantic dependencies.
2. To mitigate semantic oversmoothing and cross-scale interference prevalent in hierarchical decoding, we design a Multi-Scale Adaptive Fusion Module (MAFM). This module dynamically recalibrates multi-stage features, thereby preserving edge connectivity and topological consistency of narrow roads during feature fusion.
3. To enhance robustness against shadow-induced occlusions and background noise characterized by similar textures, we introduce a Dual-Spectrum Aggregation Module

(DualSpec). Employing FFT-based spatial–frequency decoupling, it independently refines phase and amplitude spectra before adaptively fusing them with spatial features, thereby significantly enhancing structural integrity.

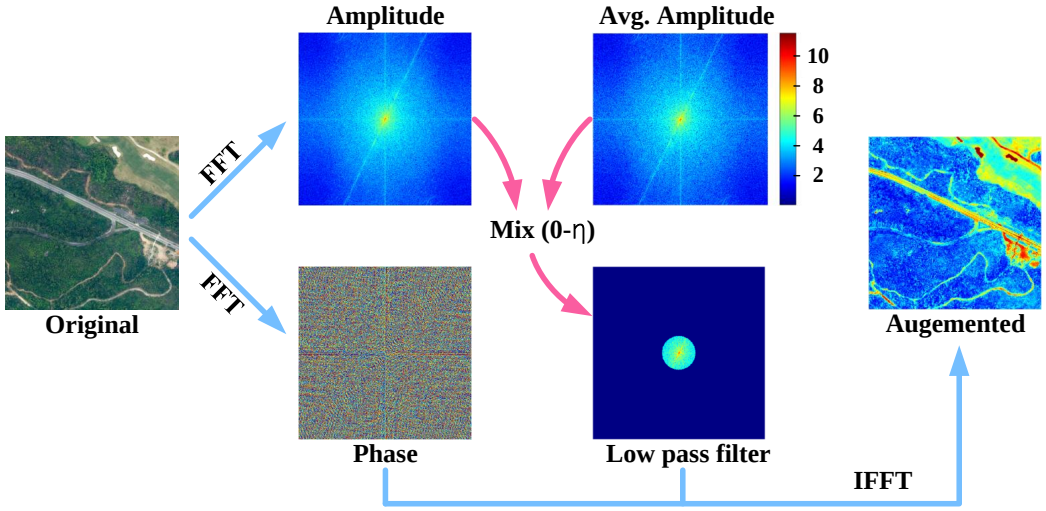


Figure 2. Illustration of the FFT applied to a road image. The phase spectrum retains essential semantic structures, including road shapes and boundaries, whereas the amplitude spectrum governs texture distribution and background appearance.

2. Related Work

2.1. CNN- and Transformer-Based Road Extraction Methods

Early CNN-based methods focused on enhancing boundary delineation and multi-scale feature representation [31,32]. UNet [4] employs an encoder–decoder structure with skip connections to fuse semantic and spatial features, but its limited depth restricts performance on fine-scale regions. Zhou et al. [33] proposed D-LinkNet by integrating skip connections with a multi-scale feature extraction mechanism. Li et al. [34] introduced BD-LinkNetPlus as an improvement upon D-LinkNet. Yi et al. [35] utilized multiple parallel branches with varying dilation rates and global average pooling (GAP) to capture global information. Wang et al. [36] introduced FRCFNet, which fuses multi-scale features for road extraction. Hua et al. [37] proposed MADSNet by integrating strip convolutions and attention mechanisms to improve topological continuity. Nevertheless, the inherent locality of convolutions restricts long-range dependency modeling, frequently causing topological fractures under occlusions or illumination variations. Although CNN-based methods have made significant progress, they still face challenges when dealing with complex remote-sensing scenes. Their limited receptive fields restrict the ability to capture global contextual information, making it difficult to mine crucial spatial dependencies between geo-objects and accurately segment remote-sensing imagery.

In recent years, ViTs, which treat 2D images as 1D sequences for modeling, have gradually become mainstream for computer vision tasks, especially semantic segmentation [6,38]. ViTs circumvent receptive field limitations by modeling global context through self-attention across image patches. Luo et al. [39] proposed BDTNet, incorporating pyramid or deformable attention mechanisms to capture multi-scale contextual information. Yang et al. [40] introduced OARENet, which adopts varying dilation rates to capture texture features in semantic feature maps. Furthermore, Yu et al. [41] proposed SegRoadv2, a hybrid network that strategically fuses convolutional branches with Transformers to preserve fine-grained spatial details and maintain road connectivity structures. Despite these advancements, ViT-based road extraction faces dual bottlenecks: (1) contextual in-

formation alone is insufficient for disambiguating roads in scenes characterized by low inter-class variance and high intra-class variability; (2) the quadratic computational complexity of self-attention severely restricts the scalability of these models for high-resolution remote-sensing imagery.

2.2. Application of Mamba Architectures in Deep Learning

Mamba, an SSM-based architecture, has recently garnered significant attention in vision tasks due to its ability to model long-range dependencies while maintaining linear computational complexity, offering a favorable alternative to Transformer-based models. VMamba [16] introduces an orthogonal SSM into a linear-complexity visual backbone, effectively bridging one-dimensional sequence modeling with two-dimensional spatial structures.

In the field of remote sensing, several Mamba-based semantic segmentation frameworks have been developed. RS-Mamba [17] pioneered the application of SSMs for remote-sensing semantic segmentation, while CM-UNet [19] synergizes CNN-based local feature extraction with Mamba-based global context modeling. Chen et al. [18] proposed RS^{3}Mamba, which explores Mamba-Transformer hybrids for remote-sensing semantic segmentation. Building upon this, architectures such as TDFNet [42] further investigate Mamba-CNN hybrids to enhance spatial dependency modeling. Furthermore, Wang et al. [43] introduced Pyramid-Mamba, which employs a dense spatial pyramid pooling (DSPP) module to encode multi-scale semantic features and a pyramid fusion Mamba (PFM) module to reduce semantic redundancy during feature fusion. However, akin to ViTs, pure Mamba architectures remain deficient in extracting fine-grained localized spatial details. Integrating Anisotropic Directional Attention into a Mamba-based backbone for localized feature refinement presents a promising paradigm to reconcile linear computational efficiency with robust topological preservation in complex scenes.

2.3. Application of Frequency-Domain Information in Deep Learning

The aforementioned methods predominantly extract road features within the spatial domain. However, due to variations in satellite viewing angles, VHR remote-sensing imagery frequently exhibits significant color and texture similarities between roads and surrounding buildings. Relying solely on spatial-domain features renders the extraction process susceptible to interference from spectrally similar backgrounds, thereby inducing errors in road topology. To address this, recent studies have explored frequency-domain representations. For instance, Zhou et al. [44] integrated spatial and frequency branches via modality-aware Fourier transforms to capture global context, while FDINet [45] and FIMP [29] employ frequency-decoupled interactions and adaptive filtering to isolate subtle, task-specific variations. Yu et al. [46] introduced FMISeg, demonstrating the efficacy of spectral–visual interactions in guided segmentation tasks. Furthermore, Song et al. [47] recently proposed a cross-domain Mamba network that jointly learns spatial and frequency representations for robust SAR oil spill detection, showcasing the potential of integrating Mamba with frequency-domain modeling for complex remote-sensing tasks. Concurrently, Xue et al. [48] introduced SDFFNet, which employs a dual-branch architecture to decouple spatial and frequency features, enhancing the model's capacity to capture both local and global information.

Despite their proven success in medical imaging and change detection, the explicit integration of frequency-domain modeling in remote-sensing road extraction remains underexplored. Specifically, the synergistic fusion of spatial structural features and frequency-domain global context to resolve topological ambiguities caused by shadow occlusions and complex backgrounds remains a critical, unresolved gap.

3. Methodology

This section details the overall network architecture and the design principles of its key components. Mathematical formulations are provided to facilitate a clear understanding of the proposed method.

3.1. Overall Architecture

The overall architecture of CoastMamba is illustrated in Figure 3a. The model follows an encoder–decoder paradigm and comprises three components: an ADAMamba-based encoder, an MAFM, and a DualSpec decoder. The encoder leverages ADA to capture local connectivity and Mamba blocks to model long-range dependencies. MAFM adaptively fuses multi-scale features to preserve narrow-road details, while the DualSpec decoder employs FFT-based spatial–frequency decoupling to robustly reconstruct roads under shadowed and complex backgrounds.

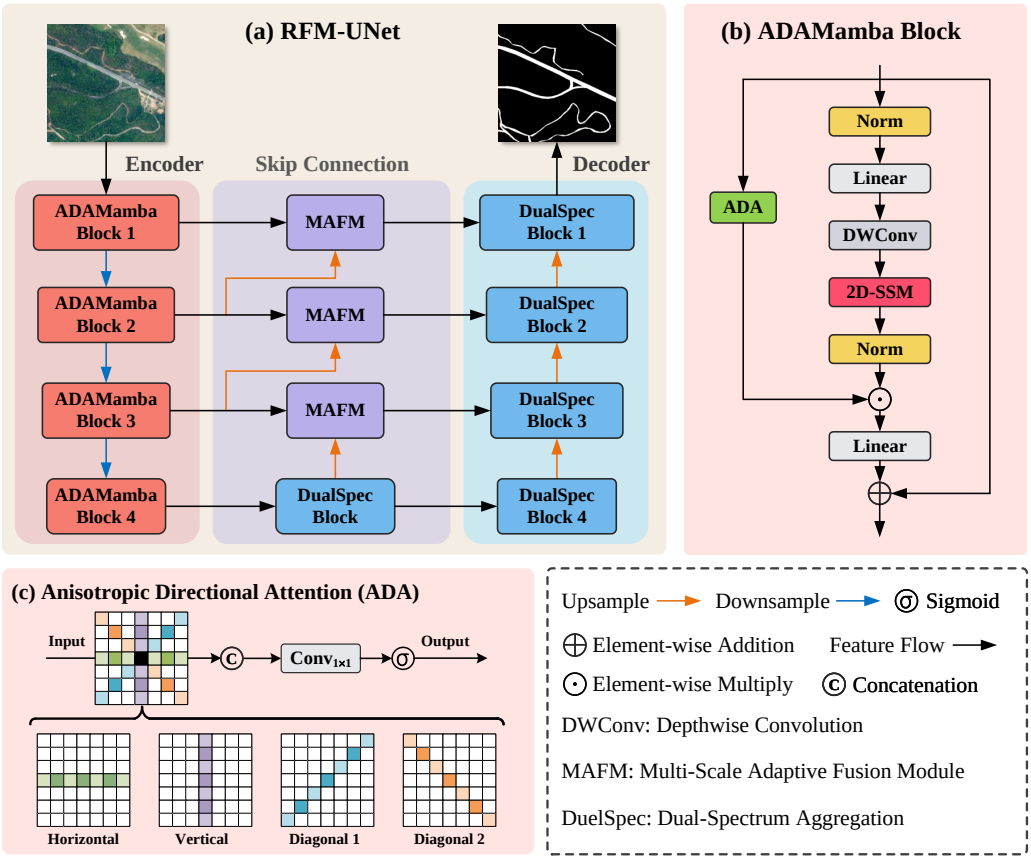


Figure 3. Overview of the proposed RFM-UNet architecture. The framework consists of three main components: an ADAMamba-based encoder integrating Anisotropic Directional Attention with Mamba blocks, a MAFM for adaptive feature aggregation, and a DualSpec-based decoder utilizing FFT-based spatial–frequency decoupling.

3.2. ADAMamba Block

Large-scale road extraction from VHR imagery necessitates models that effectively integrate global semantic information with local structural features. While ViT-based architectures have demonstrated remarkable success in computer vision via self-attention mechanisms, their quadratic computational complexity with respect to input size limits scalability for high-resolution remote-sensing data. To address this limitation, we propose the ADAMamba block, as illustrated in Figure 3b. This architecture integrates ADA with Mamba to preserve the global modeling capability of ViTs while efficiently capturing local features. Unlike traditional isotropic convolutions, which introduce substantial background

noise when expanding the receptive field, the ADA module decouples spatial encoding into four lightweight directional branches. This design facilitates the precise capture of long-range linear dependencies in road networks across multi-dimensional spatial orientations.

To capture anisotropic topological features, the input feature map $X \in \mathbb{R}^{C\times H\times W}$ first undergoes channel compression via a 1×1 convolution to constrain the computational cost of the multi-branch parallel structure. This yields a reduced feature map:

$$X_{red} = \text{SiLU}(\text{BN}(W_{reduce} \otimes X)), \tag{1}$$

where $\otimes$ denotes the matrix multiplication; $\text{BN}(\cdot)$ and $\text{SiLU}(\cdot)$ denote batch normalization and the SiLU activation function, respectively; $W_{reduce} \in \mathbb{R}^{d\times C\times 1\times 1}$ represents the convolution weights; and $d = C/r$ signifies the reduced channel dimension. Subsequently, X_{red} is processed by four parallel depthwise convolution ($DWConv$) branches to extract contextual information along the horizontal (H), vertical (V), main diagonal (D_1), and anti-diagonal (D_2) orientations.

For horizontal and vertical extraction, standard asymmetric convolutions are employed:

$$Y_H = DWConv_{1\times 3}(X_{red}), \quad Y_V = DWConv_{3\times 1}(X_{red}). \tag{2}$$

For diagonal branches, 3×3 depthwise convolution kernels are masked to strictly constrain the receptive field. Let $W_{3\times 3}$ denote the convolution weights; non-diagonal elements are nullified via the Hadamard product with mask matrices $M_{main}, M_{anti} \in \mathbb{R}^{3\times 3}$, which contain values of 1 on the main and anti-diagonals, respectively, and 0 elsewhere:

$$Y_{D_1} = (W_{3\times 3} \odot M_{main}) \otimes X_{red}, \quad Y_{D_2} = (W_{3\times 3} \odot M_{anti}) \otimes X_{red}. \tag{3}$$

where $\odot$ denotes the Hadamard product. Each branch utilizes an asymmetric zero-padding strategy to maintain spatial alignment. The resulting directional features are concatenated along the channel dimension:

$$Y_{concat} = [Y_H, Y_V, Y_{D_1}, Y_{D_2}], \quad Y_{concat} \in \mathbb{R}^{4d\times H\times W}. \tag{4}$$

To mitigate feature isolation between orientations and align channel dimensions, a 1×1 fusion convolution is applied, followed by a sigmoid activation to generate the final anisotropic attention map:

$$A_{dir} = \sigma(\text{BN}(W_{fusion} \otimes Y_{concat})), \tag{5}$$

where $\sigma(\cdot)$ is the sigmoid function and $W_{fusion} \in \mathbb{R}^{C\times 4d\times 1\times 1}$ represents the fusion weights. Notably, the batch normalization parameters in $\text{BN}(\cdot)$ are initialized to 0; this forces A_{dir} to approach 0.5 initially, thereby enhancing training stability.

Concurrently, the Mamba branch processes the input feature X to model long-range spatial dependencies via a lightweight 2D Selective Scan Module (2D-SSM). The 2D-SSM flattens image features into 1D sequences and scans them along four directions: top-left to bottom-right, bottom-right to top-left, top-right to bottom-left, and bottom-left to top-right. This multi-directional scanning captures long-range dependencies through a selective state–space model, after which the sequences are merged to reconstruct the 2D structure.

The Mamba block comprises a linear projection, depthwise convolution (DWConv), 2D-SSM, layer normalization (LN), and a final linear layer. The forward propagation is formulated as:

$$F_1 = \text{LN}(\text{2D-SSM}(\text{SiLU}(\text{DWConv}(\text{Linear}(X))))), \tag{6}$$

$$F_{out} = \text{Linear}(F_1 \odot A_{dir}) + X. \tag{7}$$

Here, the output of the Mamba branch, F_1, is modulated by the attention map A_{dir} via element-wise multiplication ($\odot$). This mechanism adaptively highlights continuous tubular structures while suppressing background noise, ensuring stable gradient flow. Finally, a residual connection integrates the original input X to produce the block's output, F_{out}. This architecture effectively balances global context modeling with local structural preservation, proving particularly suitable for VHR road extraction. The pseudocode is provided in Algorithm 1.

Algorithm 1 ADAMamba Block

Require: Input feature $X \in \mathbb{R}^{B\times C\times H\times W}$, Reduction ratio r
Ensure: Output feature $F_{\text{out}} \in \mathbb{R}^{B\times C\times H\times W}$
1. Anisotropic Directional Attention (ADA) Extraction
$X_{\text{red}} \leftarrow \text{SiLU}(\text{BN}(\text{Conv}_{1\times1}(X)))$ {Channel compression to $d = C/r$}
$Y_H \leftarrow \text{DWConv}_{1\times3}(X_{\text{red}})$ {Horizontal context}
$Y_V \leftarrow \text{DWConv}_{3\times1}(X_{\text{red}})$ {Vertical context}
$Y_{D_1} \leftarrow (W_{3\times3} \odot M_{\text{main}}) \otimes X_{\text{red}}$ {Main-diagonal context}
$Y_{D_2} \leftarrow (W_{3\times3} \odot M_{\text{anti}}) \otimes X_{\text{red}}$ {Anti-diagonal context}
$Y_{\text{concat}} \leftarrow \text{Concat}([Y_H, Y_V, Y_{D_1}, Y_{D_2}], \text{dim} = 1)$ {$Y_{\text{concat}} \in \mathbb{R}^{B\times4d\times H\times W}$}
$A_{\text{dir}} \leftarrow \sigma(\text{BN}_{\gamma=0}(\text{Conv}_{1\times1}(Y_{\text{concat}})))$ {BN initialized to 0 for stability}
2. 2D Selective State–Space Modeling (Mamba Branch)
$X_{\text{proj}} \leftarrow \text{Linear}(X)$
$X_{\text{dw}} \leftarrow \text{DWConv}(X_{\text{proj}})$
$X_{\text{ssm}} \leftarrow \text{2D-SSM}(\text{SiLU}(X_{\text{dw}}))$ {Multi-directional sequence scan}
$F_1 \leftarrow \text{LN}(X_{\text{ssm}})$
3. Modulation and Residual Fusion
$F_{\text{mod}} \leftarrow F_1 \odot A_{\text{dir}}$ {Adaptive structural modulation}
$F_{\text{out}} \leftarrow \text{Linear}(F_{\text{mod}}) + X$ {Residual connection}
return F_{out}

3.3. Multi-Scale Adaptive Fusion Module (MAFM)

In encoder–decoder architectures, directly summing or concatenating features from shallow, high-resolution branches with those from deep, low-resolution branches often causes fine-grained geometric details to be dominated by high-level semantic features. This imbalance can result in inaccurate road boundaries or topological disconnections in narrow areas, as fine structural details are easily disrupted.

To address this limitation, we propose the MAFM as shown in Figure 4a. Unlike simple feature aggregation, MAFM introduces a cross-attention mechanism where shallow features actively guide and modulate deep representations before fusion. By strictly aligning tensor dimensions and adaptively weighting the features, MAFM effectively prevents deep semantics from overwhelming shallow geometric cues, thereby mitigating the "dominance" problem and preserving narrow-road connectivity.

Let $F_s \in \mathbb{R}^{C_s\times H_s\times W_s}$ and $F_d \in \mathbb{R}^{C_d\times H_d\times W_d}$ denote the input shallow and deep features, respectively. To generate the shallow guidance signal, we apply global max pooling and average pooling to F_s in parallel. The pooled descriptors are processed by a shared channel squeeze-and-excitation (cSE) block to extract channel-wise statistics. We then apply a 1×1 convolution $\text{Conv}_{s\to d}$ to align the channel dimension from C_s to C_d, followed by a sigmoid activation to produce the shallow attention map $A_s \in \mathbb{R}^{C_d\times1\times1}$:

$$A_s = \sigma\Big(\text{Conv}_{s\to d}\Big(\text{cSE}_{\text{shared}}(\text{MaxPool}(F_s)) + \text{cSE}_{\text{shared}}(\text{AvgPool}(F_s))\Big)\Big). \tag{8}$$

Simultaneously, the deep features F_d are refined through a spatial and channel squeeze-and-excitation (scSE) block to capture internal spatial–channel dependencies. A subsequent depthwise separable convolution (DSConv) and sigmoid activation yield the deep attention map $A_d \in \mathbb{R}^{C_d \times H_d \times W_d}$:

$$A_d = \sigma(\text{DSConv}(\text{scSE}(F_d))). \tag{9}$$

The core of MAFM lies in the dual-modulation of the deep features. We apply both the shallow geometric guidance A_s (spatially broadcasted) and the deep semantic attention A_d to modulate F_d, obtaining the refined deep representation F_d^{att}:

$$F_d^{\text{att}} = F_d \odot A_s \odot A_d, \tag{10}$$

To perform the final fusion, F_d^{att} is spatially aligned with the shallow features via bilinear interpolation to match the resolution (H_s, W_s), yielding F_d^{up}. The features are concatenated along the channel dimension and passed through a fusion block (comprising a 1×1 convolution, batch normalization, ReLU, and a 3×3 DSConv) to integrate the multi-scale representations. A residual connection from F_d^{up}, aligned to the output channel dimension C_{out} via Conv_{res}, is added to yield the final fused feature $F_f \in \mathbb{R}^{C_{\text{out}} \times H_s \times W_s}$:

$$F_f = \text{FusionConv}\Big(\text{Concat}(F_s, F_d^{\text{up}})\Big) + \text{Conv}_{\text{res}}(F_d^{\text{up}}). \tag{11}$$

This explicitly aligned, cross-attention-driven fusion guarantees that critical spatial cues from the shallow branch are structurally preserved during semantic integration. The pseudocode for MAFM is provided in Algorithm 2.

Algorithm 2 Multi-Scale Adaptive Fusion Module (MAFM)

Require: Shallow features $F_s \in \mathbb{R}^{B \times C_s \times H_s \times W_s}$, Deep features $F_d \in \mathbb{R}^{B \times C_d \times H_d \times W_d}$
Ensure: Fused features $F_f \in \mathbb{R}^{B \times C_{\text{out}} \times H_s \times W_s}$

1. Shallow Feature Guidance Generation
$F_s^{\text{avg}} \leftarrow \text{AdaptiveAvgPool2d}(F_s)$
$F_s^{\text{max}} \leftarrow \text{AdaptiveMaxPool2d}(F_s)$
$A_s^{\text{raw}} \leftarrow \text{cSE}_{\text{shared}}(F_s^{\text{max}}) + \text{cSE}_{\text{shared}}(F_s^{\text{avg}})$
$A_s^{\text{proj}} \leftarrow \text{Conv}_{s \to d}(A_s^{\text{raw}})$ {Align channels: $C_s \to C_d$}
$A_s \leftarrow \text{Sigmoid}(A_s^{\text{proj}})$
2. Deep Feature Self-Refinement
$F_d^{\text{scse}} \leftarrow \text{scSE_block}(F_d)$
$F_d^{\text{dsconv}} \leftarrow \text{DSConv}(F_d^{\text{scse}})$
$A_d \leftarrow \text{Sigmoid}(F_d^{\text{dsconv}})$
3. Cross-Attention Modulation
$F_d^{\text{att}} \leftarrow F_d \odot \text{ExpandAs}(A_s, F_d)$ {Spatial broadcast of A_s}
$F_d^{\text{att}} \leftarrow F_d^{\text{att}} \odot A_d$ {Element-wise modulation}
4. Spatial Alignment
$F_d^{\text{up}} \leftarrow \text{Interpolate}(F_d^{\text{att}}, \text{size} = (H_s, W_s))$ {Bilinear upsampling}
5. Feature Fusion
$F_{\text{concat}} \leftarrow \text{Concat}([F_s, F_d^{\text{up}}], \text{dim} = 1)$
$F_{\text{refined}} \leftarrow \text{FusionConv}(F_{\text{concat}})$
$F_f \leftarrow F_{\text{refined}} + \text{Conv}_{\text{res}}(F_d^{\text{up}})$
return F_f

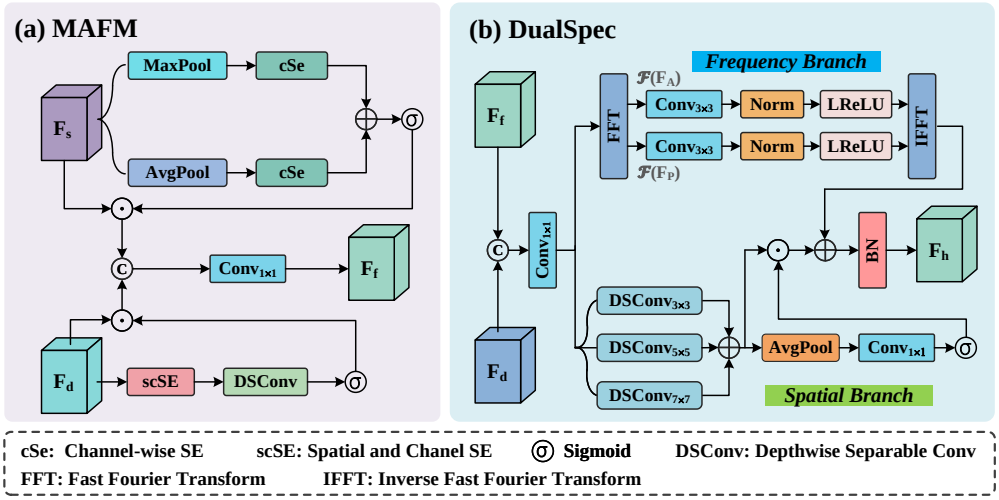


Figure 4. Detailed design of the MAFM and DualSpec. The MAFM adaptively fuses multi-scale features from the encoder and decoder branches, while the DualSpec decouples spatial features into frequency-domain representations and adaptively fuses them to enhance robustness under shadow occlusions.

3.4. Dual-Spectrum Aggregation Module (DualSpec)

In remote-sensing road extraction, complex backgrounds and uneven illumination often lead to blurred or aliased spatial-domain features under shadow occlusions, thereby reducing road identification accuracy. Frequency-domain transformations explicitly decompose spectral information, effectively separating global structural patterns captured by the phase spectrum from fine-grained details represented by the amplitude spectrum, such as road edges and intersections.

To fully exploit this property, we propose a novel DualSpec decoder as shown in Figure 4b. Unlike conventional feature aggregators, DualSpec explicitly decouples the complex frequency spectrum into independent real-valued amplitude and phase components. This allows standard convolutional layers to independently optimize high-frequency details and low-frequency structural shapes before recombining them with spatial-domain features.

Specifically, given the decoder feature F_{dec} and the encoder fusion feature F_{enc}, we first align their spatial resolutions via bilinear interpolation and concatenate them. A bottleneck layer (comprising a 1×1 convolution, batch normalization, and ReLU) reduces the channel dimension to generate the intermediate hybrid feature $X \in \mathbb{R}^{C \times H \times W}$:

$$X = \text{ReLU}\Big(\text{BN}\big(\text{Conv}_{1\times1}(\text{Concat}(F^{\uparrow}_{\text{dec}}, F_{\text{enc}}))\big)\Big). \tag{12}$$

The feature X is then processed by two parallel branches: the Frequency Branch and the Spatial Branch.

Frequency Branch: To enable cross-domain cooperative optimization, we map X to the frequency domain using the 2D Real FFT (RFFT2), yielding the complex spectrum $\mathcal{X}$. Instead of applying operations directly to the complex tensor, we decouple $\mathcal{X}$ into its real-valued amplitude A and phase Φ:

$$\mathcal{X} = \text{RFFT2}(X), \tag{13}$$

$$A = |\mathcal{X}|, \quad \Phi = \angle\mathcal{X}. \tag{14}$$

To independently enhance these components, we apply channel-wise 1×1 grouped convolutions ($\text{Conv}_{\text{group}}$), followed by batch normalization and Leaky ReLU activations. This creates two distinct optimization paths, yielding the refined amplitude A' and phase Φ':

$$A' = \mathrm{LReLU}\Big(\mathrm{BN}\big(\mathrm{Conv}_{\mathrm{group}}(A)\big)\Big), \tag{15}$$

$$\Phi' = \mathrm{LReLU}\Big(\mathrm{BN}\big(\mathrm{Conv}_{\mathrm{group}}(\Phi)\big)\Big). \tag{16}$$

The complex spectrum is then reconstructed using Euler's formula and transformed back to the spatial domain via the 2D Inverse Real FFT (IRFFT2) to produce the frequency-enhanced feature F_{freq}:

$$\mathcal{X}' = A' \odot \Big(\cos(\Phi') + j\sin(\Phi')\Big), \tag{17}$$

$$F_{\mathrm{freq}} = \mathrm{IRFFT2}(\mathcal{X}'). \tag{18}$$

Spatial Branch: Simultaneously, the spatial branch captures multi-scale contextual information using parallel DSConv with varying kernel sizes ($k \in \{3, 5, 7\}$). The multi-scale features are summed to form F_{s1}:

$$F_{\mathrm{s1}} = \sum_{k\in\{3,5,7\}} \mathrm{DSConv}_{k\times k}(X). \tag{19}$$

To emphasize critical spatial regions, an attention mechanism is applied. Global spatial dependencies are captured via adaptive average pooling, compressed by a 1×1 convolution, and activated by a sigmoid function. The modulated features are further refined by a final 3×3 DSConv and batch normalization to yield F_{spatial}:

$$F_{\mathrm{s2}} = F_{\mathrm{s1}} \odot \sigma\Big(\mathrm{Conv}_{1\times1}\big(\mathrm{AvgPool}(F_{\mathrm{s1}})\big)\Big), \tag{20}$$

$$F_{\mathrm{spatial}} = \mathrm{BN}\Big(\mathrm{DSConv}_{3\times3}(F_{\mathrm{s2}})\Big). \tag{21}$$

Hybrid Fusion: Finally, the outputs from both branches are aggregated to leverage the complementary strengths of the structural frequency cues and the multi-scale spatial details. The final fused representation F_{out} is obtained by element-wise addition followed by batch normalization:

$$F_{\mathrm{out}} = \mathrm{BN}(F_{\mathrm{freq}} + F_{\mathrm{spatial}}). \tag{22}$$

This explicitly decoupled dual-spectrum mechanism significantly improves the model's robustness and representational capacity, ensuring accurate road extraction even under severe shadow occlusions.

4. Experimental Settings and Results

4.1. Datasets

To comprehensively evaluate the effectiveness and generalization capability of the proposed RFM-UNet for remote-sensing road extraction, we conducted extensive experiments on three representative, publicly available datasets: Massachusetts Roads [49], DeepGlobe Road [50], and CHN6-CUG Roads [51]. Table 1 summarizes the statistics of the original images in the datasets, and Figure 5 illustrates their characteristics.

1. Massachusetts Roads: This dataset comprises 1171 high-resolution aerial images from the state of Massachusetts, each with a size of 1500×1500 pixels, covering approximately 2.25 km^2 per image and over 2600 km^2 in total. The road structures exhibit significant variations in width, shape, texture, and color, making the dataset highly challenging. Following the official data split, all images were padded to

1536 × 1536 pixels and cropped into 512 × 512 patches, resulting in 9972 training images, 126 validation images, and 441 test images.

2. DeepGlobe Road: Released for the CVPR 2018 DeepGlobe Road Extraction Challenge, this dataset contains 8570 high-resolution satellite images from regions such as India, Indonesia, and Thailand. Each image has a size of 1024 × 1024 pixels with a spatial resolution of 0.5 m, covering diverse landscapes from urban to rural areas. Since official test labels were not publicly available, we utilized 6226 labeled images and split them into 4696 training samples and 1530 testing samples for our experiments. All images were directly cropped into 512 × 512 patches for training and evaluation.
3. CHN6-CUG Roads: Developed by the URSmart team at the China University of Geosciences (Wuhan), this VHR imagery dataset contains pixel-level manual annotations. It covers six representative urban areas: Beijing, Shanghai, Wuhan, Shenzhen, Hong Kong, and Macau, capturing diverse urbanization levels and city characteristics. Road annotations include both covered and uncovered roads, categorized according to coverage extent, and encompass various types such as railways, highways, urban roads, and rural roads. The dataset consists of 4511 images of size 512 × 512 pixels, with 3608 used for training and 903 reserved for testing and evaluation.

Table 1. Summary of three public road datasets used in the experiments.

Dataset	GSD (m)/Bands	Patch Size	Train	Val/Test	Region	Release
Massachusetts	1/RGB	1500 × 1500	1108	14/49	Massachusetts	2013
DeepGlobe	0.5/RGB	1024 × 1024	4696	-/1530	South and Southeast Asia	2018
CHN6-CUG	0.5/RGB	512 × 512	3608	-/903	China	2021

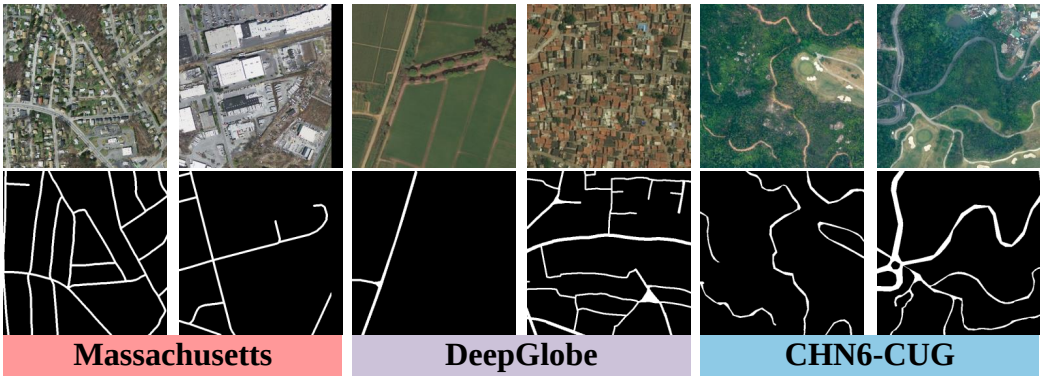


Figure 5. Examples of images and masks from each dataset used in the experiments.

4.2. Evaluation Metrics

Following standard practice in semantic segmentation evaluation, four pixel-level metrics are adopted, all derived from the confusion matrix: Precision (P), Recall (R), F1-score ($F1$), and Intersection over Union (IoU). These are formulated as:

$$P = \frac{TP}{TP + FP} \tag{23}$$

$$R = \frac{TP}{TP + FN} \tag{24}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{25}$$

$$\text{IoU} = \frac{TP}{TP + FP + FN} \tag{26}$$

where TP, FP, and FN denote the numbers of true positives, false positives, and false negatives, respectively. Among these, $F1$ and IoU serve as the primary composite indicators of overall segmentation performance.

Furthermore, to assess the model's ability to preserve road topology, we employ road-specific topological indicators: Centerline Dice (clDice) and Average Path Length Similarity (APLS). The clDice metric evaluates the overlap between predicted and ground-truth road centerlines, while APLS measures the similarity of path lengths between the predicted and reference road networks, providing a comprehensive assessment of topological accuracy.

$$\text{clDice} = \frac{2 \times |C_{pred} \cap C_{gt}|}{|C_{pred}| + |C_{gt}|} \tag{27}$$

$$\text{APLS} = 1 - \frac{1}{N} \sum_{i=1}^{N} \min\left(\frac{|L_{pred}(i) - L_{gt}(i)|}{L_{gt}(i)}, 1 \right) \tag{28}$$

where C_{pred} and C_{gt} represent the predicted and ground-truth centerline sets, respectively, and $L_{pred}(i)$ and $L_{gt}(i)$ denote the lengths of the i-th path in the predicted and ground-truth road networks. APLS values closer to 1 indicate better topological similarity.

4.3. Experimental Setup

The experimental environment is detailed in Table 2. All models were implemented using the PyTorch framework and trained on a single NVIDIA RTX 4090 GPU with 24 GB of VRAM. The AdamW optimizer was employed with a cosine annealing learning rate schedule. Data augmentation included random horizontal and vertical flipping to enhance model generalization. For the Massachusetts Roads dataset, all models were trained from scratch for 105 epochs with an initial learning rate of 1×10^{-3} and a batch size of 12. For the DeepGlobe and CHN6-CUG datasets, all models were fine-tuned using weights pre-trained on the Massachusetts dataset, leveraging a unified transfer learning strategy, for 105 epochs with a learning rate of 5×10^{-4} and a batch size of 8. During inference, test-time augmentation (TTA), including horizontal and vertical flips, was applied to all models to improve prediction stability and robustness. All experiments were conducted under identical conditions to ensure fair comparisons.

Table 2. Default configurations for the experiments.

Parameter	Value	Description
Language	Python 3.12	Programming language used
Framework	PyTorch 2.4.0	Deep-learning framework
GPU	NVIDIA RTX 4090	Graphics processing unit used for training
Random Seed	42	Seed for reproducibility
Optimizer	AdamW	Optimization algorithm
Learning Rate	1×10^{-3}	Initial learning rate (Massachusetts)
Batch Size	12	Number of samples per batch (Massachusetts)
Epochs	105	Number of training epochs

4.4. Comparative Experiments with Different Image Sizes

To determine the optimal input scale, we evaluated RFM-UNet on the Massachusetts dataset (1 m GSD) using four patch sizes: 128 × 128, 256 × 256, 512 × 512, and 1024 × 1024 pixels. Although larger inputs theoretically encapsulate broader spatial contexts, Table 3 demonstrates that the 512 × 512 configuration yields the optimal performance (IoU = 68.59%, clDice = 89.13%, and APLS = 74.04%), notably outperforming the 1024 × 1024 configuration. This indicates that a moderate receptive field optimally balances fine-grained detail preservation with contextual aggregation for the proposed architecture. Consequently, a

512 × 512 input size was adopted for all subsequent experiments to maximize both accuracy and computational efficiency.

Table 3. Performance comparison with different input image sizes on the Massachusetts dataset.

Patch Size	P (%)	R (%)	F1 (%)	IoU (%)	cIDice (%)	APLS (%)
128 × 128	79.20	78.82	79.01	65.30	86.98	69.58
256 × 256	80.56	78.75	79.65	66.18	87.84	72.62
512 × 512	**81.51**	**81.24**	**81.37**	**68.59**	**89.13**	**74.04**
1024 × 1024	81.20	81.13	81.15	68.27	89.09	73.71

Note: The best results are highlighted in bold.

4.5. Experimental Results

To further validate the effectiveness of the proposed method, we conducted comparative experiments on three benchmark datasets: Massachusetts, DeepGlobe, and CHN6-CUG. We compared RFM-UNet with several representative methods, including UNet [4], LinkNet [52], D-LinkNet [33], DeepLabV3+ [53], TransUNet [54], SwinUNet [55], UNet* (based on ConvNeXt), RCFSNet [56], RoadExNet [57], OARENet [40], RS^{3}Mamba [58], SegRoadv2 [41], and MADSNet [37]. All methods were trained and evaluated under identical experimental settings, including the same transfer learning strategy for the DeepGlobe and CHN6-CUG datasets.

While the full metrics are detailed in Table 4, we highlight the primary performance here. On the Massachusetts Roads dataset, RFM-UNet achieves an IoU score of 68.59%, outperforming the second-best method by 0.77%. On the DeepGlobe Road dataset, it obtains an IoU of 70.71%, surpassing the second-best method by 1.45%. Similarly, on the CHN6-CUG Roads dataset, RFM-UNet reaches an IoU of 64.52%, yielding an improvement of 0.19%. Compared with state-of-the-art methods, RFM-UNet consistently demonstrates significant improvements across all datasets, proving its superior overall performance. These results validate the strong representational capacity and excellent generalization ability of the proposed model in complex remote-sensing scenarios characterized by diverse road geometries and structural patterns across different regions.

Table 4. Quantitative comparison with popular methods on three public road datasets (%).

Method	Year	Massachusetts Roads				DeepGlobe Road				CHN6-CUG Roads			
		P	R	F1	IoU	P	R	F1	IoU	P	R	F1	IoU
UNet	2015	77.29	72.13	74.62	59.51	78.29	70.92	74.42	59.26	70.67	71.66	71.16	55.23
LinkNet	2017	79.18	73.71	75.52	61.74	75.62	73.75	74.67	59.85	74.92	71.43	73.13	57.64
D-LinkNet	2018	77.88	74.52	76.16	61.50	77.01	76.27	76.64	62.12	76.28	71.89	74.01	58.75
DeepLabV3+	2018	75.47	77.97	76.70	62.25	78.20	76.24	77.20	62.87	75.21	74.59	74.90	59.87
TransUNet	2021	78.37	77.29	77.82	63.70	77.80	77.02	77.41	63.14	74.32	77.02	75.64	60.83
SwinUNet	2022	78.12	77.07	77.59	63.59	78.44	77.30	77.87	63.76	74.43	76.47	75.44	60.56
UNet*	2022	79.81	78.08	78.93	65.19	78.63	77.82	78.22	64.23	74.81	76.79	75.78	61.01
RCFSNet	2022	79.23	77.62	78.43	64.49	78.85	79.35	79.10	65.43	75.65	77.29	76.46	61.89
RoadExNet	2023	80.38	75.98	78.18	64.17	76.94	79.42	78.16	64.15	75.23	77.30	76.25	61.61
OARENet	2024	77.48	79.92	78.70	64.88	81.72	81.95	81.84	69.26	77.83	72.43	75.03	60.04
RS-Mamba	2024	**86.27**	76.03	80.82	67.82	**83.14**	80.00	81.54	68.84	74.63	**79.77**	77.11	62.75
RS^{3}Mamba	2024	80.87	77.65	79.26	65.65	80.45	79.30	79.87	66.49	77.82	72.80	75.23	60.29
SegRoadv2	2025	79.84	79.02	79.43	65.87	81.06	80.98	81.02	68.09	72.10	78.76	75.28	60.36
MADSNet	2025	79.40	**81.36**	80.38	67.18	80.58	80.79	80.69	67.63	77.49	79.11	78.29	64.33
RFM-UNet		81.51	81.24	**81.37**	**68.59**	81.77	**83.95**	**82.84**	**70.71**	**80.07**	76.87	**78.44**	**64.52**

Note: The best results are highlighted in bold, and the second-best results are underlined.

Figure 6 presents a visual comparison of different methods on the Massachusetts Roads dataset. The first two columns display the input images and their corresponding ground-truth annotations, while the remaining six columns show the predictions from RFM-UNet, MADSNet, RS^{3}Mamba, RS-Mamba, OARENet, and RoadExNet. Overall, most models accurately detect prominent features such as intersections and straight road segments, demonstrating their ability to capture structural patterns. However, challenges remain in handling curved roads and shadowed areas caused by buildings or vegetation, which often lead to misclassification between road patterns and background textures. In contrast, RFM-UNet exhibits clear advantages in distinguishing road pixels from visually similar non-road categories (e.g., trees and buildings), thereby reducing false positives. It also maintains superior continuity in road structures and effectively mitigates the impact of shadows and occlusions. Furthermore, RFM-UNet accurately delineates road boundaries under varying illumination conditions, demonstrating robust performance across diverse imaging scenarios. These improvements are attributed to the proposed ADAMamba architecture, which efficiently integrates global contextual information with local spatial details to enhance structural consistency.

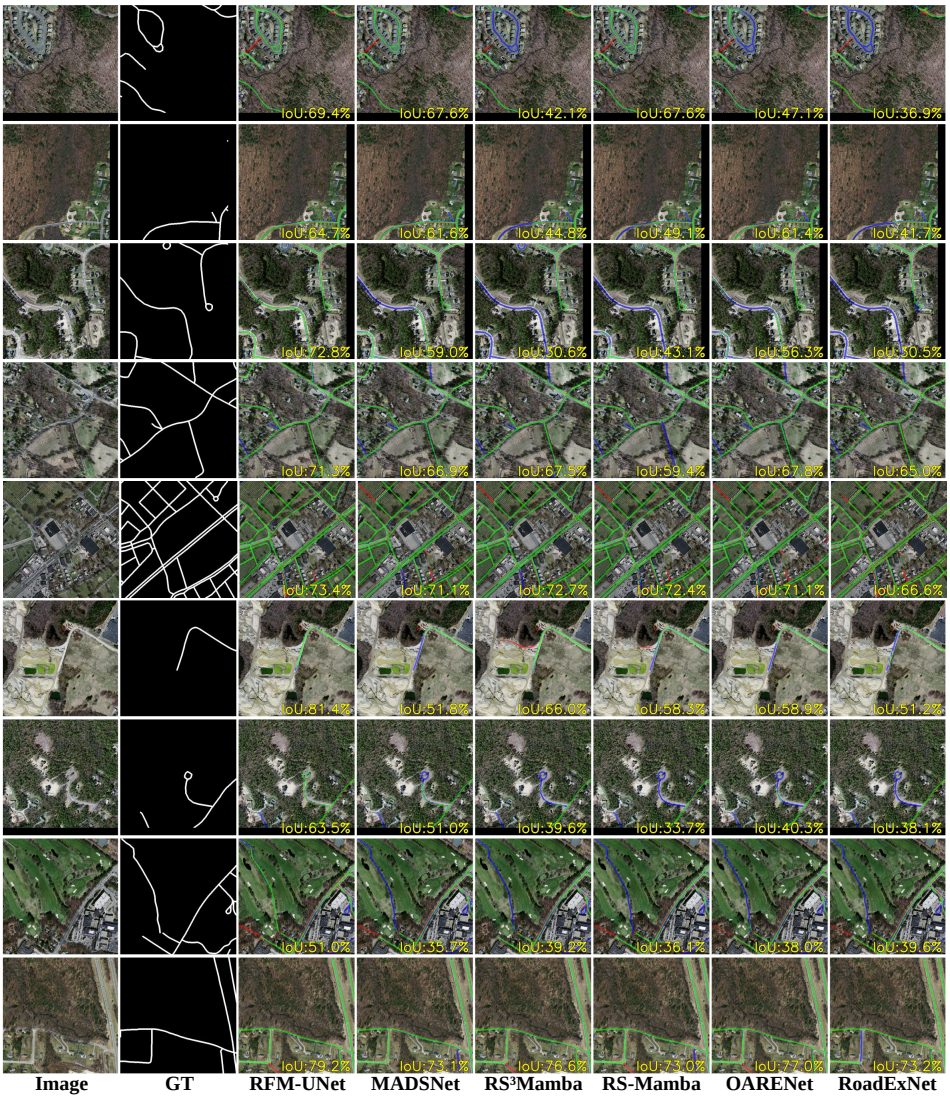


Figure 6. Visual comparison of RFM-UNet, MADSNet, RS^{3}Mamba, RS-Mamba, OARENet, and RoadExNet on the Massachusetts Roads dataset. True positives are marked in green, false positives in red, and false negatives in blue. The numbers in the corners of the predicted images represent the corresponding patch-level IoU scores.

Figure 7 shows qualitative results on the DeepGlobe dataset. In complex rural and suburban environments, road extraction is challenged by dense vegetation, occlusions, and illumination variations. As observed in areas such as intersections, narrow roads, and farmland-adjacent regions, most models misclassify non-road elements as roads due to a reliance on local pixel information, which limits their ability to capture global semantic relationships between roads and their surroundings. In contrast, RFM-UNet maintains strong structural continuity and topological consistency even when roads are partially obscured or visually confounded with background textures. Moreover, the model effectively distinguishes non-road regions resembling dirt roads, demonstrating superior robustness and generalization in complex rural and suburban scenes.



Figure 7. Visual comparison of RFM-UNet, MADSNet, RS^{3}Mamba, RS-Mamba, OARENet, and RoadExNet on the DeepGlobe Road dataset. True positives are marked in green, false positives in red, and false negatives in blue. The numbers in the corners of the predicted images represent the corresponding patch-level IoU scores.

Figure 8 presents qualitative results on the CHN6-CUG Roads dataset. We evaluated model performance under various shadow-interference scenarios, including continuous shadows and vegetation occlusions, as well as complex conditions such as intersections

and curved roads. While most models perform adequately in non-shadowed regions, their performance degrades significantly in densely shaded areas due to a reliance on local pixel features. This limitation hinders the capture of structural relationships, particularly the connectivity required along curved roads. RFM-UNet, in contrast, demonstrates superior performance under these challenging conditions by effectively leveraging spatial–frequency features for robust and stable predictions. These results further confirm the advantage of the proposed model in jointly integrating spatial and frequency-domain representations to improve road extraction in complex remote-sensing scenarios.

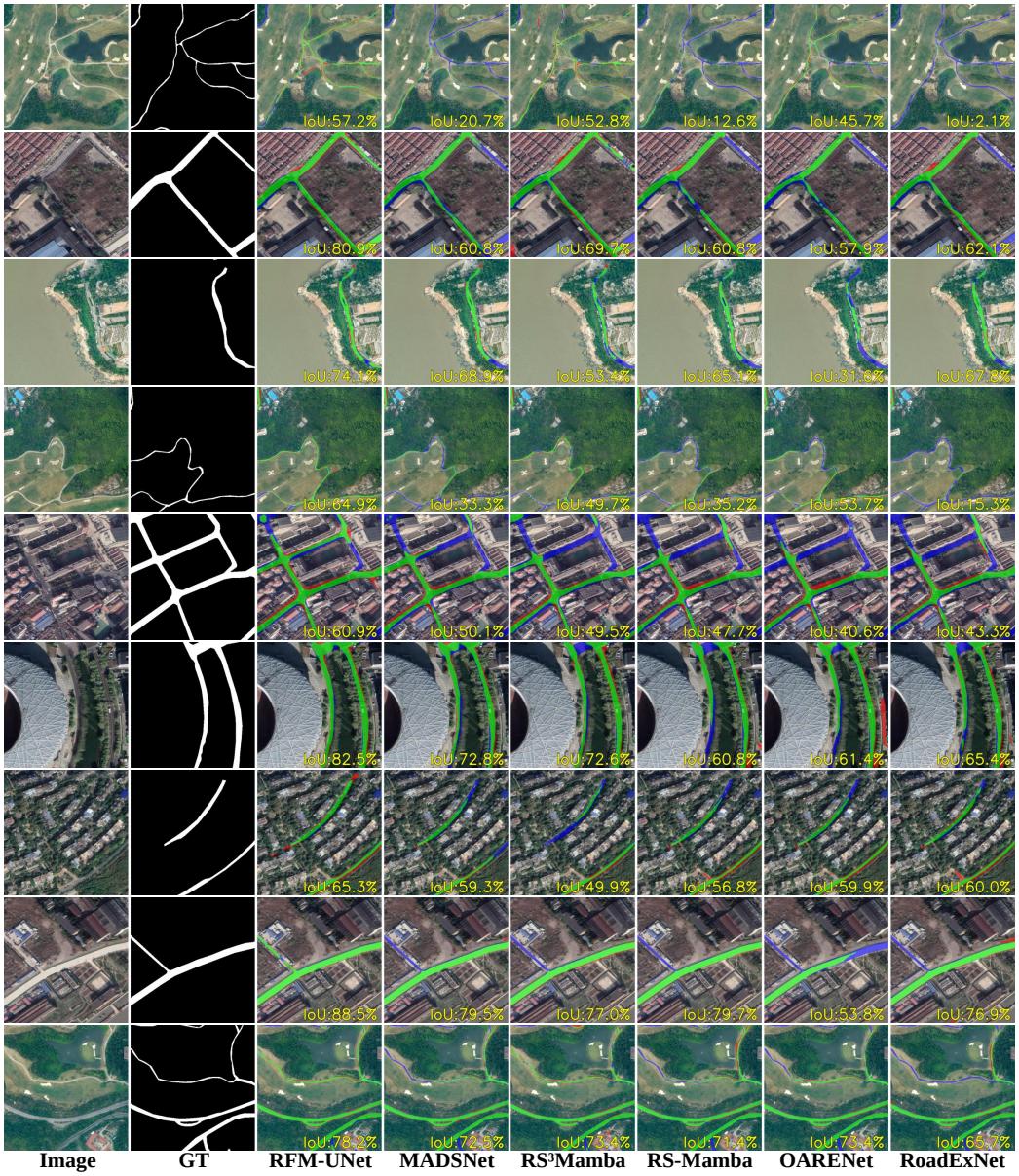


Figure 8. Visual comparison of RFM-UNet, MADSNet, RS^{3}Mamba, RS-Mamba, OARENet, and RoadExNet on the CHN6-CUG Roads dataset. True positives are marked in green, false positives in red, and false negatives in blue. The numbers in the corners of the predicted images represent the corresponding patch-level IoU scores.

4.6. Experimental Analysis of Road Topology Accuracy

To further assess the topological accuracy of the extracted road networks, we conducted a detailed analysis using the cIDice and APLS metrics. Unlike traditional pixel-level metrics, these topological metrics are crucial for evaluating whether the extracted roads

form a continuous, graph-level structure, which is essential for practical routing applications. Table 5 presents the comparative topological results across the three datasets.

Overall, RFM-UNet consistently achieves the highest cIDice and APLS scores compared to other state-of-the-art methods. Specifically, on the Massachusetts Roads dataset, RFM-UNet obtains a cIDice of 89.13% and an APLS of 74.04%, outperforming the second-best method by 0.31% and 0.84%, respectively. On the DeepGlobe Road dataset, which features complex rural and suburban landscapes prone to vegetation occlusion, our model reaches a cIDice of 90.91% and an APLS of 69.39%, outperforming the second-best method by 1.02% and 0.34%, respectively, thereby demonstrating strong robustness against frequent road breaks. Furthermore, on the highly challenging CHN6-CUG Roads dataset—characterized by dense urban canopies, shadows, and complex intersections—RFM-UNet still secures the highest scores (cIDice: 78.26%, APLS: 72.17%), outperforming the second-best method by 0.15% and 0.14%, respectively, and maintaining a consistent lead over the competitive MADSNet.

These quantitative improvements indicate the model's superior ability to preserve road connectivity and accurately represent road centerlines. As shown in Figure 9, this topological advantage can be primarily attributed to the ADAMamba encoder, which captures long-range topological dependencies to bridge spatial gaps caused by occlusions, and the MAFM module, which facilitates coherent multi-scale feature aggregation to maintain the continuity of roads with varying widths. Maintaining such topological integrity is particularly critical for downstream applications such as navigation and urban planning, where even minor discontinuities or topological errors can lead to routing failures.

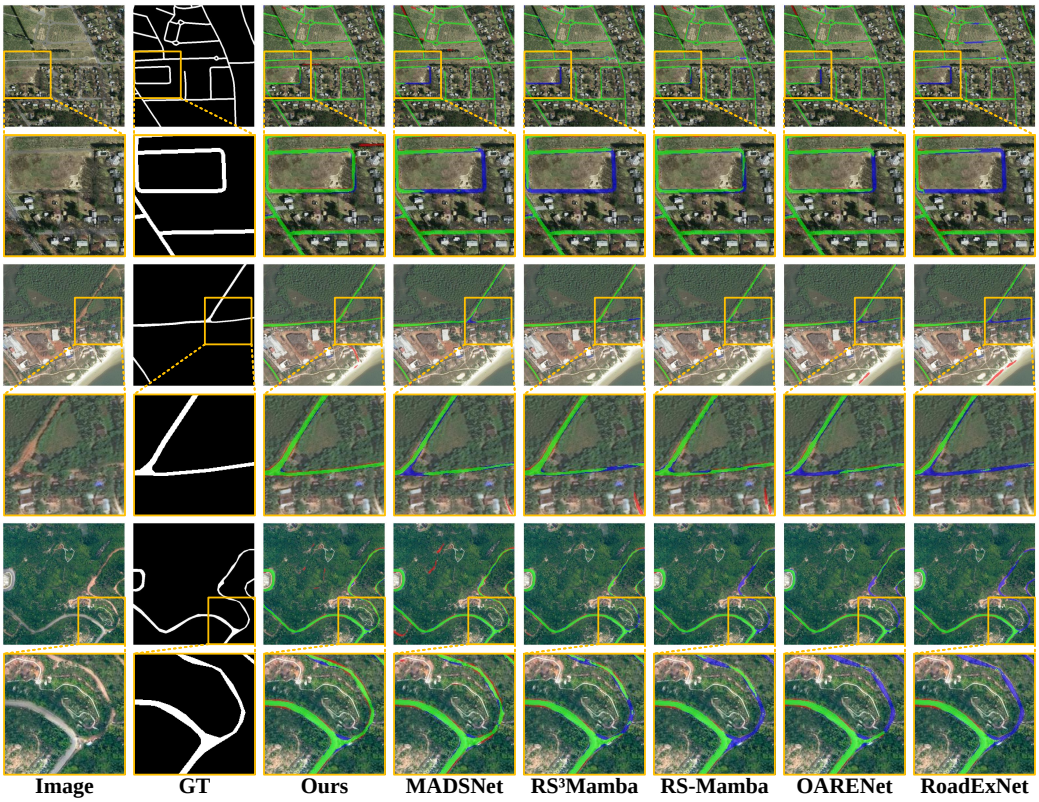


Figure 9. Topological analysis of road extraction methods on the Massachusetts, DeepGlobe, and CHN6-CUG datasets. RFM-UNet consistently achieves the highest cIDice and APLS scores, indicating superior preservation of road connectivity and accurate representation of road centerlines. The yellow regions show the local enlarged view of the road topology, highlighting the continuity and connectivity of the extracted road networks.

Table 5. Topological analysis of road extraction methods on three datasets.

Method	Massachusetts Roads		DeepGlobe Road		CHN6-CUG Roads	
	cIDice (%)	APLS (%)	cIDice (%)	APLS (%)	cIDice (%)	APLS (%)
UNet	83.98	69.12	84.62	66.45	71.66	64.50
LinkNet	84.61	69.98	84.90	67.05	73.92	65.28
D-LinkNet	84.44	68.79	85.32	67.20	74.53	65.74
DeepLabV3+	85.75	69.18	85.60	67.45	74.61	66.50
TransUNet	85.95	69.35	85.74	67.74	72.84	66.88
SwinUNet	85.90	69.26	86.12	67.82	75.06	67.10
UNet*	87.33	72.52	86.55	68.21	75.47	67.29
RCFSNet	86.85	71.27	86.89	68.17	76.65	67.89
RoadExNet	86.64	71.01	87.58	68.06	76.52	67.97
OARENet	87.09	71.69	<u>89.89</u>	<u>69.05</u>	75.53	68.04
RS-Mamba	<u>88.82</u>	<u>73.20</u>	89.15	68.92	77.63	70.15
RS^{3}Mamba	87.72	72.12	88.18	68.03	75.82	68.29
SegRoadv2	87.97	72.33	88.80	68.62	76.10	68.36
MADSNet	88.02	72.85	88.62	68.26	<u>78.11</u>	<u>72.03</u>
RFM-UNet	**89.13**	**74.04**	**90.91**	**69.39**	**78.26**	**72.17**

Note: The best results are highlighted in bold, and the second-best results are underlined.

4.7. Model Complexity and Computational Efficiency Analysis

To evaluate the computational efficiency of RFM-UNet, we compared its parameters, floating-point operations (FLOPs), memory footprint, and frames per second (FPS) with those of several popular models, as summarized in Table 6. The FLOPs and FPS were calculated based on an input image size of $3 \times 512 \times 512$ with a batch size of 1 on an NVIDIA RTX 4090 GPU.

Table 6. Comparison of model efficiency on an NVIDIA RTX 4090.

Method	Backbone	Year	Params (M)	FLOPs (G)	Memory (MB)	FPS
LinkNet	CNN	2017	21.64	23.26	186.28	97.19
D-LinkNet	CNN	2018	31.09	26.71	230.40	258.83
MADSNet	CNN	2025	29.73	32.58	449.42	92.20
RoadExNet	ViT	2023	59.20	277.40	1035.78	26.44
OARENet	ViT	2024	33.48	36.27	370.51	85.35
SegRoadv2	ViT	2025	54.99	109.42	717.61	44.90
RS-Mamba	SSM	2024	50.37	46.33	644.54	48.87
RS^{3}Mamba	SSM	2024	43.32	39.56	472.62	51.99
RFM-UNet	SSM		29.86	30.13	428.69	60.70

As shown in Table 6, RFM-UNet achieves a favorable balance between model complexity and computational cost. With 29.86 million parameters and 30.13 GFLOPs, it is significantly more efficient than many Transformer-based architectures. For instance, Transformer models like RoadExNet consume up to 1035.78 MB of memory and run at only 26.44 FPS, making them computationally prohibitive for processing large-scale remote-sensing imagery. In contrast, RFM-UNet requires only 428.69 MB of memory, which is notably lower than other state–space models (SSMs) such as RS-Mamba (644.54 MB) and RS^{3}Mamba (472.62 MB), as well as the recent CNN-based MADSNet (449.42 MB). Although pure CNN architectures like D-LinkNet exhibit extremely high inference speeds (258.83 FPS) due to their lightweight structures, they often suffer from limitations in global context modeling. RFM-UNet reaches 60.70 FPS, which is significantly faster than heavier ViT-based models (e.g., RoadExNet and SegRoadv2) and surpasses other SSMs, thereby ensuring efficient inference for high-resolution data without compromising global percep-

tion capabilities. Figure 10 further illustrates the trade-off between accuracy and efficiency, highlighting RFM-UNet's ability to achieve high segmentation performance with relatively low computational demands.

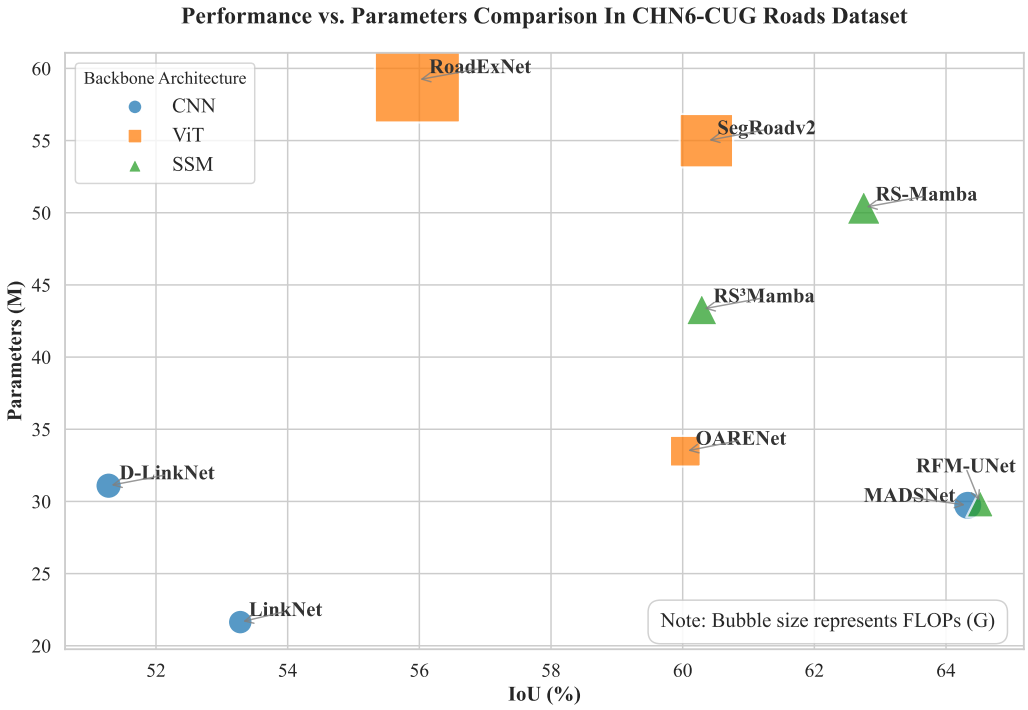


Figure 10. Comparison of model efficiency on the CHN6-CUG Roads dataset. RFM-UNet achieves a superior balance between accuracy and computational efficiency compared to other popular methods.

4.8. Ablation Study

In this section, we conducted a series of ablation experiments to validate the efficacy of the proposed RFM-UNet and to demonstrate the necessity of its various components. The ablation experiments for the overall framework were implemented on the CHN6-CUG Roads dataset, chosen for its complex road structures and diverse environmental conditions, which provide a rigorous testbed for evaluating the model's performance. The baseline model utilizes a standard VMamba backbone without the ADAMamba, MAFM, and DualSpec modules. Each component was incrementally added to the baseline to assess its individual contribution to the overall performance.

As detailed in Table 7, the incremental integration of the ADAMamba, MAFM, and DualSpec modules is assessed. Furthermore, Grad-CAM [59] was employed to visualize activation distributions (Figure 11), elucidating the internal feature response mechanisms during inference.

Table 7. Ablation study of the RFM-UNet architecture on the CHN6-CUG Roads dataset (%).

Baseline	ADAMamba	MAFM	DualSpec	Params	FLOPs	P	R	F1	IoU	cIDice	APLS
✓				23.13 M	24.09 G	74.21	74.33	74.27	59.07	73.65	65.49
✓	✓			25.55 M	26.78 G	75.42	77.53	76.47	60.54	74.49	67.75
✓		✓		26.39 M	26.92 G	74.14	78.04	76.04	61.34	75.78	69.72
✓			✓	24.20 M	24.60 G	76.44	78.49	77.45	63.20	76.43	71.86
✓	✓	✓	✓	29.86 M	30.13 G	**80.07**	**76.87**	**78.44**	**64.52**	**78.26**	**72.17**

Note: The best results are highlighted in bold.

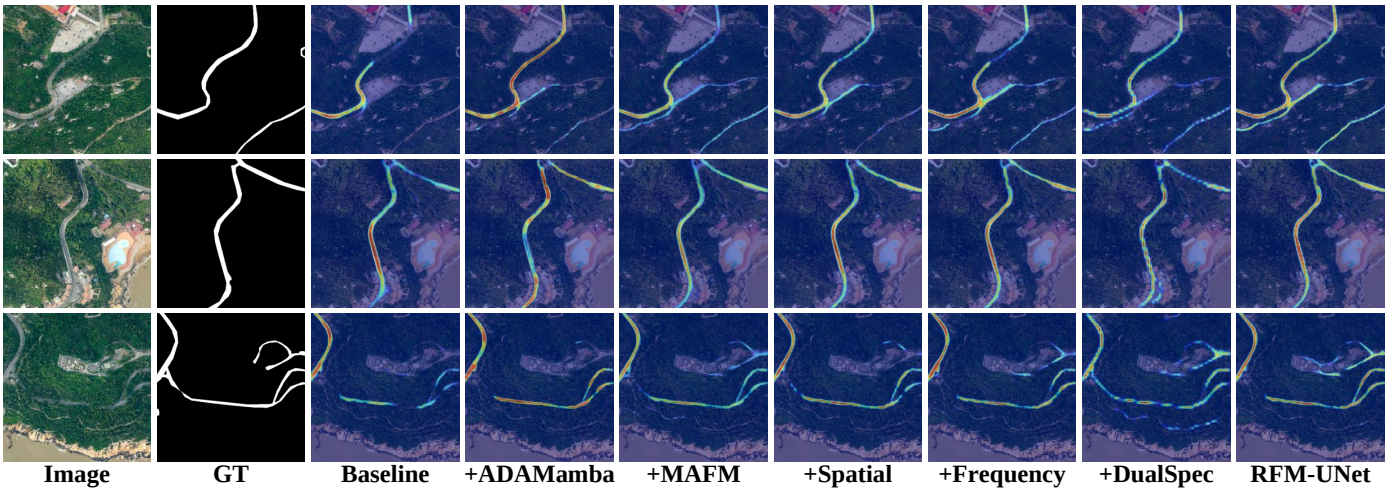


Figure 11. Grad-CAM visualization of the ablation study on the CHN6-CUG Roads dataset. The colors in the figure represent the intensity of activation, with warmer colors indicating higher activation levels.

4.8.1. Effectiveness of ADAMamba

We first investigated the impact of the ADAMamba block through an ablation study on the CHN6-CUG Roads dataset. Upgrading the baseline VMamba encoder with the ADAMamba module increases the IoU from 59.07% to 60.54%, clDice from 73.65% to 74.49%, and APLS from 65.49% to 67.75%. This indicates that the anisotropic attention mechanism successfully captures localized directional features to enhance topological continuity. As shown in Figure 11, compared with the baseline, the ADAMamba encoder effectively focuses on complex road structures and intersections, thereby improving the model's ability to maintain connectivity in challenging scenarios.

4.8.2. Effectiveness of MAFM

Subsequently, we evaluated the contribution of the MAFM. The baseline with MAFM increases the IoU from 59.07% to 61.34%, clDice from 73.65% to 75.78%, and APLS from 65.49% to 69.72%. As visualized in Figure 11, the MAFM module enhances the model's ability to capture multi-scale semantic information, particularly in regions with varying road widths and complex intersections. This results in improved delineation of road boundaries and better preservation of structural integrity.

4.8.3. Effectiveness of DualSpec

Finally, we evaluated the contribution of the DualSpec module. The baseline with DualSpec increases the IoU from 59.07% to 63.20%, clDice from 73.65% to 76.43%, and APLS from 65.49% to 71.86%. As shown in Figure 11, the DualSpec module effectively enhances the model's ability to recover spatial details and mitigate shadow-induced degradation, leading to more accurate road segmentation in occluded regions.

To further elucidate the contributions of the dual spatial–frequency representation, we conducted a comparative analysis, as shown in Table 8, among the Spatial-only variant, the Frequency-only variant, and the full DualSpec model. The Spatial-only variant achieves an IoU of 62.21%, clDice of 75.92%, and APLS of 69.33%, while the Frequency-only variant achieves an IoU of 62.62%, clDice of 76.31%, and APLS of 70.48%. In contrast, the full DualSpec model achieves an IoU of 63.20%, clDice of 76.43%, and APLS of 71.86%. Furthermore, we visualize the mean activation maps for three different regions in Figure 12. These results demonstrate that the dual-branch design is not merely a redundant addition of parameters; rather, the integration of frequency-domain features in DualSpec provides complementary

information that enhances the model's ability to capture fine-grained structural details, particularly in shadow-occluded regions.

Table 8. Comparison of Spatial-only, Frequency-only, and full DualSpec models on the CHN6-CUG Roads dataset (%).

Model	P	R	F1	IoU	cIDice	APLS
Spatial-only	74.84	**78.67**	76.71	62.21	75.92	69.33
Frequency-only	75.43	78.66	77.01	62.62	76.31	70.48
DualSpec	**76.44**	78.49	**77.45**	**63.20**	**76.43**	**71.86**

Note: The best results are highlighted in bold.

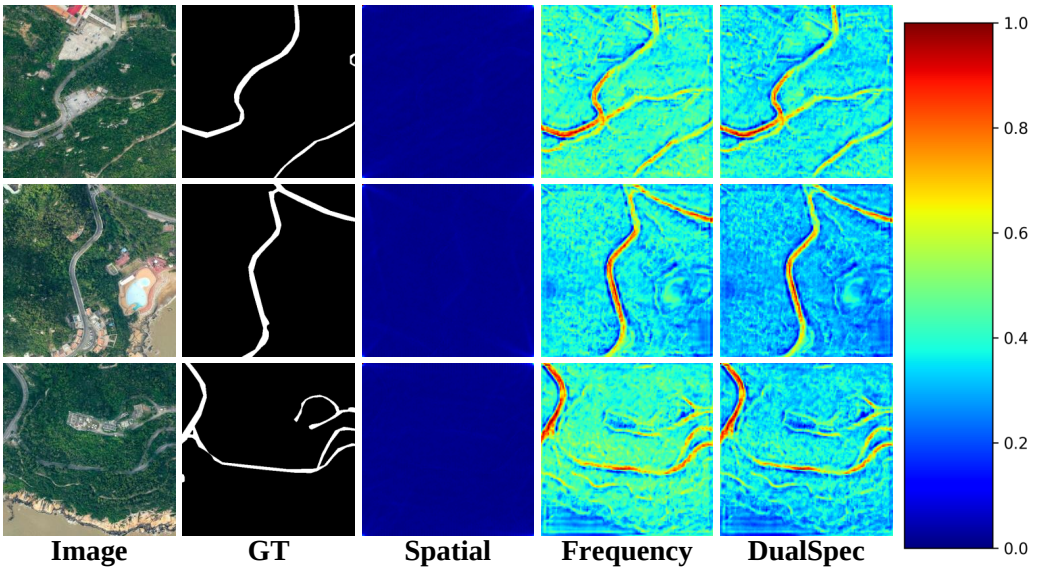


Figure 12. Mean activation maps for three different regions comparing the Spatial-only variant and the full DualSpec model. The full DualSpec model demonstrates improved feature representation, particularly in shadow-occluded areas.

5. Conclusions

This paper proposed RFM-UNet, a novel deep-learning framework designed for high-precision road extraction in VHR remote-sensing imagery. Central to this framework is the ADAMamba encoder, developed to jointly model local geometric details and global topological dependencies. Building upon this encoder, the framework integrates specialized MAFM and DualSpec modules for adaptive multi-scale feature fusion and spatial–frequency decoupling. The ADAMamba encoder effectively addresses critical challenges in VHR imagery, including complex topological structures and narrow-road discontinuities, thereby enhancing the structural integrity of road networks. Furthermore, the MAFM and DualSpec modules mitigate semantic oversmoothing and refine shadow-occluded boundaries, fully exploiting hierarchical feature information.

Extensive experiments on the Massachusetts Roads, DeepGlobe Road, and CHN6-CUG Roads datasets demonstrated that RFM-UNet consistently outperforms current state-of-the-art methods. Ablation studies further validated the efficacy and individual contributions of the ADAMamba, MAFM, and DualSpec components. Given the robust performance of the spatial–frequency synergy paradigm, RFM-UNet holds significant potential for advancing intelligent transportation and environmental monitoring. However, our method also presents certain limitations, particularly in scenarios involving extreme shadow occlusions and highly complex urban environments. Future research will focus on

enhancing the model's robustness to these challenging conditions, potentially through the integration of additional contextual information and advanced attention mechanisms.

Author Contributions: Conceptualization, P.S. and P.Y.; Methodology, P.S. and P.Y.; Writing—original draft, P.S. and P.Y.; Writing—review and editing, P.S., P.Y., X.Z., S.W., J.T., Y.H., L.Z., G.L. and M.L.; Visualization, P.S. and P.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by Fujian Provincial Natural Science Foundation of China under Grant 2025J011279; the Natural Science Foundation of Xiamen, China under Grant 3502Z202573066; the Fujian Provincial Science and Technology Innovation Project of Natural Resources under Grants KY-080000-04-2025-014 and KY-100000-04-2025-023; and the National Natural Science Foundation of China under Grant 42471378.

Data Availability Statement: The datasets used and analyzed during the current study are available from the following sources: Massachusetts road dataset: https://www.kaggle.com/datasets/balraj98/massachusetts-roads-dataset (accessed on 10 July 2026); DeepGlobe road dataset: http://deepglobe.org/challenge.html (accessed on 10 July 2026); CHN6-CUG road dataset: https://grzy.cug.edu.cn/zhuqiqi/zh_CN/index.htm (accessed on 10 July 2026). The source code is publicly available at https://github.com/FF7CA/RFMUNet (accessed on 10 July 2026).

Conflicts of Interest: Author Lujun Zhang was employed by the company Shanghai Investigation, Design and Research Institute. 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. Barzohar, M.; Cooper, D.B. Automatic finding of main roads in aerial images by using geometric-stochastic models and estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *18*, 707–721.
2. Chai, D.; Forstner, W.; Lafarge, F. Recovering line-networks in images by junction-point processes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013*; IEEE: Piscataway, NJ, USA, 2013; pp. 1894–1901.
3. Wang, D.; Li, J.; Zhu, S. Detecting urban hot regions by using massive geo-tagged image data. *Neurocomputing* **2021**, *428*, 325–331. [CrossRef]
4. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015*; Proceedings, Part III 18; Springer: Cham, Switzerland, 2015; pp. 234–241.
5. Fan, Y.; Hong, C.; Zeng, G.; Liu, L. A deep convolutional encoder–decoder–restorer architecture for image deblurring. *Neural Process. Lett.* **2024**, *56*, 27. [CrossRef]
6. Jiang, X.; Li, Y.; Jiang, T.; Xie, J.; Wu, Y.; Cai, Q.; Jiang, J.; Xu, J.; Zhang, H. RoadFormer: Pyramidal deformable vision transformers for road network extraction with remote sensing images. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *113*, 102987. [CrossRef]
7. Chen, J.; Yin, H.; Zhang, D. A self-adaptive classification method for plant disease detection using GMDH-Logistic model. *Sustain. Comput. Inform. Syst.* **2020**, *28*, 100415. [CrossRef]
8. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929. [CrossRef]
9. Weng, W.; Hou, F.; Gong, S.; Chen, F.; Lin, D. Attribute graph clustering via transformer and graph attention autoencoder. *Intell. Data Anal.* **2025**, *29*, 306–319. [CrossRef]
10. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021*; IEEE: Piscataway, NJ, USA, 2021; pp. 10012–10022.
11. Liu, L.; Li, P.; Wang, D.; Zhu, S. A wind turbine damage detection algorithm designed based on YOLOv8. *Appl. Soft Comput.* **2024**, *154*, 111364. [CrossRef]
12. Lin, R.; He, Y.; Xu, M. Method of sensitive data mining based on Pan-Bull algebra. *Wirel. Netw.* **2022**, *28*, 2733–2741.
13. Xie, Y.; Hong, C.; Zhuang, W.; Liu, L.; Li, J. HOGFormer: High-order graph convolution transformer for 3D human pose estimation. *Int. J. Mach. Learn. Cybern.* **2025**, *16*, 599–610.

14. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. In Proceedings of the First Conference on Language Modeling, Philadelphia, PA, USA, 7–9 October 2024.
15. Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; Wang, X. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv* **2024**, arXiv:2401.09417.
16. Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; Liu, Y. Vmamba: Visual state space model. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 103031–103063. [CrossRef]
17. Zhao, S.; Chen, H.; Zhang, X.; Xiao, P.; Bai, L.; Ouyang, W. Rs-mamba for large remote sensing image dense prediction. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5633314 [CrossRef]
18. Chen, K.; Chen, B.; Liu, C.; Li, W.; Zou, Z.; Shi, Z. Rsmamba: Remote sensing image classification with state space model. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 8002605. [CrossRef]
19. Liu, M.; Dan, J.; Lu, Z.; Yu, Y.; Li, Y.; Li, X. CM-UNet: Hybrid CNN-Mamba UNet for remote sensing image semantic segmentation. *arXiv* **2024**, arXiv:2405.10530.
20. Zheng, Y.; Zheng, W.; Du, X. Paddy-YOLO: An accurate method for rice pest detection. *Comput. Electron. Agric.* **2025**, *238*, 110777. [CrossRef]
21. Yu, S.; Wu, Y.; Li, W.; Song, Z.; Zeng, W. A model for fine-grained vehicle classification based on deep learning. *Neurocomputing* **2017**, *257*, 97–103. [CrossRef]
22. Shunzhi, Z.; Lizhao, L.; Si, C. Image feature detection algorithm based on the spread of hessian source. *Multimed. Syst.* **2017**, *23*, 105–117.
23. Chen, Y.; Feng, S.; Zhao, C.; Su, N.; Li, W.; Tao, R.; Ren, J. High-resolution remote sensing image change detection based on Fourier feature interaction and multi-scale perception. *IEEE Trans. Geosci. Remote Sens.* **2024**, *21*, 8002605.
24. Xu, C.; Ye, Z.; Mei, L.; Yu, H.; Liu, J.; Yalikun, Y.; Jin, S.; Liu, S.; Yang, W.; Lei, C. Hybrid attention-aware transformer network collaborative multiscale feature alignment for building change detection. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 5012914. [CrossRef]
25. Zhu, X.; Hu, H.; Lin, S.; Dai, J. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019*; IEEE: Piscataway, NJ, USA, 2019; pp. 9308–9316.
26. Lee, S.; Bae, J.; Kim, H.Y. Decompose, adjust, compose: Effective normalization by playing with frequency for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023*; IEEE: Piscataway, NJ, USA, 2023; pp. 11776–11785.
27. Cao, B.; Wang, Q.; Zhu, P.; Hu, Q.; Ren, D.; Zuo, W.; Gao, X. Multi-view knowledge ensemble with frequency consistency for cross-domain face translation. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *35*, 9728–9742.
28. Zhang, G.; Zhang, W.; Li, W.; Wang, L.; Cui, H. A dynamic attention mechanism for object detection in road or strip environments. *Vis. Comput.* **2025**, *41*, 4171–4181.
29. Xiong, F.; Li, T.; Yang, Y.; Zhou, J.; Lu, J.; Qian, Y. Wavelet siamese network with semi-supervised domain adaptation for remote sensing image change detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5633613. [CrossRef]
30. Yi, J.; Han, W.; Lai, F. YOLOv8n-DDSW: An efficient fish target detection network for dense underwater scenes. *PeerJ Comput. Sci.* **2025**, *11*, e2798. [CrossRef]
31. Mnih, V.; Hinton, G.E. Learning to detect roads in high-resolution aerial images. In *Proceedings of the Computer Vision—ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, 5–11 September 2010*; Proceedings, Part VI 11; Springer: Cham, Switzerland, 2010; pp. 210–223
32. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015*; IEEE: Piscataway, NJ, USA, 2015; pp. 3431–3440.
33. Zhou, L.; Zhang, C.; Wu, M. D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018*; IEEE: Piscataway, NJ, USA, 2018; pp. 182–186.
34. Li, Y.; Peng, B.; He, L.; Fan, K.; Li, Z.; Tong, L. Road extraction from unmanned aerial vehicle remote sensing images based on improved neural networks. *Sensors* **2019**, *19*, 4115. [CrossRef] [PubMed]
35. Yi, J.; Shen, Z.; Chen, F.; Zhao, Y.; Xiao, S.; Zhou, W. A lightweight multiscale feature fusion network for remote sensing object counting. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5902113. [CrossRef]
36. Wang, H.; Bai, L.; Xue, D.; Momi, M.C.; Ye, Z.; Quan, S. FRCFNet: Feature reassembly and context information fusion network for road extraction. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 2502805. [CrossRef]
37. Hua, Z.T.; Chen, S.B.; Lu, W.; Tang, J.; Luo, B. Multi-Scale Adaptive Decoder and Diverse Selection for Road Extraction in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2025**, *62*, 4411713.
38. Li, R.; Zheng, S.; Zhang, C.; Duan, C.; Wang, L.; Atkinson, P.M. ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of Fine-Resolution remotely sensed imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *181*, 84–98. [CrossRef]

39. Luo, L.; Wang, J.X.; Chen, S.B.; Tang, J.; Luo, B. BDTNet: Road extraction by bi-direction transformer from remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 2505605. [CrossRef]
40. Yang, R.; Zhong, Y.; Liu, Y.; Lu, X.; Zhang, L. Occlusion-aware road extraction network for high-resolution remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5619316. [CrossRef]
41. Yu, Z.; Chen, Z.; Xiao, K.; Lei, X.; Tang, R.; He, Q.; Sun, Z.; 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]
42. Wang, W.; Yu, P.; Li, M.; Zhong, X.; He, Y.; Su, H.; Zhou, Y. TDFNet: Twice decoding V-Mamba-CNN Fusion features for building extraction. *Geo-Spat. Inf. Sci.* **2026**, *29*, 19–38.
43. Wang, L.; Li, D.; Dong, S.; Meng, X.; Zhang, X.; Hong, D. PyramidMamba: Rethinking pyramid feature fusion with selective space state model for semantic segmentation of remote sensing imagery. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *144*, 104884. [CrossRef]
44. Zhou, M.; Huang, J.; Yan, K.; Hong, D.; Jia, X.; Chanussot, J.; Li, C. A general spatial-frequency learning framework for multimodal image fusion. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *47*, 5281–5298. [CrossRef]
45. Tang, Y.; Feng, S.; Zhao, C.; Fan, Y.; Shi, Q.; Li, W.; Tao, R. An object fine-grained change detection method based on frequency decoupling interaction for high-resolution remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *62*, 5600213.
46. Yu, B.; Yang, J.; Du, Z.; Huang, Y.; Li, C.; Wang, L. Frequency-domain Multi-modal Fusion for Language-guided Medical Image Segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Daejeon, South Korea, 23–27 September 2025*; Springer: Cham, Switzerland, 2025; pp. 278–288.
47. Song, P.; Yu, P.; Zhong, X.; Wang, L.; He, Y.; Li, G.; Qi, Z.; Su, H. A Cross-Domain Mamba Network with joint spatial-frequency learning for robust SAR oil spill detection. *Mar. Pollut. Bull.* **2026**, *229*, 119729. [CrossRef]
48. Xue, D.; Bai, L.; Zhou, W.; Zhang, Y.; Gozho, A.; Quan, S. SDFFNet: A Spatial-Frequency Domain Feature Fusion Network for Road Extraction From Very High-Resolution Satellite Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2026**, *19*, 18207–18223. [CrossRef]
49. Mnih, V. Machine Learning for Aerial Image Labeling. Ph.D. Thesis, University of Toronto, Toronto, ON, Canada, 2013.
50. Demir, I.; Koperski, K.; Lindenbaum, D.; Pang, G.; Huang, J.; Basu, S.; Hughes, F.; Tuia, D.; Raskar, R. Deepglobe 2018: A challenge to parse the earth through satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018*; IEEE: Piscataway, NJ, USA, 2018; pp. 172–181.
51. Zhu, Q.; Zhang, Y.; Wang, L.; Zhong, Y.; Guan, Q.; Lu, X.; Zhang, L.; Li, D. A global context-aware and batch-independent network for road extraction from VHR satellite imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *175*, 353–365. [CrossRef]
52. Chaurasia, A.; Culurciello, E. Linknet: Exploiting Encoder Representations for Efficient Semantic Segmentation. In *Proceedings of the 2017 IEEE Visual Communications and Image Processing (VCIP), St. Petersburg, FL, USA, 10–13 December 2017*; IEEE: Piscataway, NJ, USA, 2017; pp. 1–4.
53. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018*; Springer: Cham, Switzerland, 2018; pp. 801–818.
54. 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]
55. 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 *Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022*; Springer: Cham, Switzerland, 2022; pp. 205–218.
56. Yang, Z.; Zhou, D.; Yang, Y.; Zhang, J.; Chen, Z. Road extraction from satellite imagery by road context and full-stage feature. *IEEE Geosci. Remote Sens. Lett.* **2022**, *20*, 8000405. [CrossRef]
57. 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]
58. Ma, X.; Zhang, X.; Pun, M.O. Rs 3 mamba: Visual state space model for remote sensing image semantic segmentation. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 6011405. [CrossRef]
59. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017*; IEEE: Piscataway, NJ, USA, 2017; pp. 618–626.

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.