Acta Informatica Pragensia
2027, Volume 16, Issue 1, Forthcoming article
https://doi.org/10.18267/j.aip.323

PRAGUE UNIVERSITY OF ECONOMICS AND BUSINESS

Article Open Access

RoadNeXt-CSS: ConvNeXt-Adapted Encoder-Decoder Road Extraction for Robust Spatial Information Systems

Rachmawan Atmaji Perdana [1], **Aniati Murni Arymurthy** [2], **Risnandar Risnandar** [1]

[1] Research Center for Artificial Intelligence and Cybersecurity, Institute of Electronics and Informatics, National Research and Innovation Agency, Bandung, Indonesia
[2] Department of Computer Science, Faculty of Computer Science, University of Indonesia, Depok, Indonesia

Corresponding author: Rachmawan Atmaji Perdana (rach020@brin.go.id)

Editorial Record

First submission received:
April 21, 2026

Revision received:
June 17, 2026

Accepted for publication:
July 8, 2026

Academic Editor:
Bartosz Marcinkowski
University of Gdańsk, Poland

This article was accepted for publication by the Academic Editor upon evaluation of the reviewers' comments.

How to cite this article:
Perdana, R. A., Arymurthy, A. M., & Risnandar, R. (2027). RoadNeXt-CSS: ConvNeXt-Adapted Encoder-Decoder Road Extraction for Robust Spatial Information Systems. *Acta Informatica Pragensia*, 16(1), Forthcoming article. https://doi.org/10.18267/j.aip.323

Copyright:
© 2027 by the author(s). Licensee Prague University of Economics and Business, Czech Republic. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution License (CC BY 4.0).

Abstract

Background: Road extraction from remote sensing imagery is essential for geospatial information systems but faces challenges of thin structures, severe class imbalance and similarity between roads and non-roads.

Objective: We propose RoadNeXt-CSS, integrating a ConvNeXt encoder, inverse frequency weighting (IFW) loss and a scale-sensitive decoder to address these challenges.

Methods: RoadNeXt-CSS was trained and evaluated on the Massachusetts Roads and DeepGlobe benchmarks using intersection over union (IoU) as the primary metric. ConvNeXt extracts rich spatial-semantic features, IFW mitigates foreground-background imbalance and the scale-sensitive decoder adaptively weights multi-level features. Component-wise ablation experiments were conducted to isolate the contribution of each proposed component and paired t-tests were applied to assess statistical significance against baseline models.

Results: RoadNeXt-CSS achieves IoU of 0.6779 (Massachusetts Roads) and 0.6984 (DeepGlobe), outperforming RCFSNet and other baselines ($p < 0.001$). Ablation analysis confirms that the ConvNeXt encoder, IFW loss and scale-sensitive decoder contribute independently to performance gains.

Conclusion: The three components together provide robust and statistically significant road extraction improvements, supporting transportation planning, disaster response and navigation systems.

Index Terms

Road extraction; Remote sensing imagery; ConvNeXt; Semantic segmentation; Class imbalance.

1 INTRODUCTION

Remote sensing technology has advanced substantially since its early adoption in the 1970s, driven by improvements in satellite sensors, imaging capabilities and the availability of open geospatial datasets. These developments have enabled large-scale acquisition of high-resolution imagery and supported widespread adoption of automated image interpretation techniques. Among the most important applications is road extraction, which aims to segment road surfaces from surrounding land cover at the pixel level (Mo et al., 2024). Accurate road maps are essential for transportation planning, navigation, urban monitoring, logistics, disaster response and autonomous driving. Consequently, improving road extraction contributes not only to geospatial analysis but also to the reliability, accessibility and effectiveness of digital services that depend on spatial information (R. Liu et al., 2024).

Road extraction remains challenging despite the distinctive geometric, topological and textural characteristics of roads. Roads typically exhibit elongated structures, consistent widths and connected network patterns with intersections (W. Wang et al., 2016). However, their appearance is strongly influenced by environmental and contextual occlusions from vegetation, buildings and vehicles, variations in illumination and shadows and visual similarities between roads and other paved surfaces that frequently complicate the extraction process (D. Wei et al., 2023). These challenges become particularly pronounced in complex urban and rural environments, where roads appear at multiple scales and exhibit diverse visual characteristics.

To address these challenges, this study proposes **RoadNeXt-CSS**, a road extraction framework that integrates a modern convolutional backbone, inverse-frequency-weighted optimization and scale-sensitive feature refinement. The proposed approach aims to improve feature representation, mitigate class imbalance and enhance the extraction of road structures across varying spatial scales. A review of related work is presented in Section 1.1, followed by the identification of research gaps and contributions in Section 1.2.

1.1 Literature review

Early road extraction methods relied on conventional image processing and machine learning techniques. Thresholding followed by morphological opening was used to refine road network representations (C. Zhang et al., 1999). Probabilistic classification approaches, such as naive Bayes, were applied to estimate pixel membership probabilities and assign each pixel to the most likely class (J. Zhou et al., 2006). Active contour models were explored by Anil & Natarajan (2010), where median filtering was applied before snake-based elastic deformation starting from manually selected seed points. Multiclass support vector machines (SVMs) were also employed for segmentation and classification of multispectral imagery (Simler, 2011).

The first convolutional neural network (CNN)–based approach for road extraction was proposed by Y. Li et al. (2017), which operated on image patches. Although effective, this method incurred high computational costs and showed limited performance on high-resolution imagery. Subsequent developments led to fully convolutional networks and encoder–decoder architectures. Among these, U-Net (Ronneberger et al., 2015) became the most widely adopted, alongside SegNet, PSPNet, LinkNet (Chaurasia & Culurciello, 2017) and D-LinkNet (L. Zhou et al., 2018). Numerous variants of U-Net have since been proposed by modifying encoder backbones, decoder structures, bridge modules, skip connections and optimization strategies. An improvement to U-Net by using denser skip connection and an asymmetric convolution block was proposed in MACU-Net (R. Li et al., 2022). The transition of encoder-decoder frameworks for road extraction has increasingly incorporated transformer-based components. CAFormer (X. Wang et al., 2025), for instance, uses the dynamic routing mechanism of the BiFormer block in both the encoder and the decoder.

Attention mechanisms have been increasingly integrated to enhance contextual understanding. DRC-Net (D. Wei et al., 2023) employs the convolutional block attention module (CBAM) (Woo et al., 2018) to jointly model spatial and channel attention. Criss-cross channel attention (CCA), which captures long-range dependencies in horizontal and vertical directions, is utilized in both DRC-Net and Fe-LinkNet (Q. Wang et al., 2022). Attention gates are also applied during feature fusion between consecutive levels to suppress irrelevant responses (Soni et al., 2025). Dai et al. (2023) combined spatial deformable convolution with spatial attention to adaptively capture the topological structure of roads.

A notable innovation is CoANet (Mei et al., 2021), which introduced the strip convolution module (SCM) to capture contextual information along elongated directions. SCM was also adopted in AGF-Net (Y. Zhang et al., 2023) for multilayer feature fusion. Related approaches exploit strip pooling to preserve road continuity, as seen in MSPFE-Net (Z. Wei & Zhang, 2023), where strip pooling is applied within skip connections, and in MECA-Net (Jie et al., 2022), where SCM functions as a spatial attention mechanism combining enhanced skip features with decoder outputs.

Several studies have focused on fusing shallow spatial features with deep semantic features to improve segmentation robustness (Akhtarmanesh et al., 2024). The gating and refinement unit was proposed by Z. Zhang et al. (2022) and H. Wang et al. (2023) to eliminate ambiguous information from shallow layers. SAN (Huan et al., 2022) employs channel-based attention to integrate features from different depths, whereas RCFSNet (Yang et al., 2023) concatenates features from all encoder layers.

The bridge layer, which contains the richest semantic information, is commonly used to address occlusion and connectivity issues. Nonlocal attention mechanisms were applied at this stage by Du et al. (2023) and X. Wang et al. (2023). Q. Wang et al. (2022) further modified D-LinkNet with the DP-block, while RCFSNet enhanced global context modelling using multiscale convolution branches with vertical and horizontal kernels.

The importance of encoder selection is demonstrated in DeepRoadNet (Ahmed et al., 2024), where a state-of-the-art backbone significantly improves mIoU. While segmentation is typically derived from the final decoder layer, several studies (Tan et al., 2021; Yang et al., 2023; Z. Zhang et al., 2022) have aggregated features from multiple decoder stages to improve accuracy and convergence. The scale-sensitive neural network (Tan et al., 2021) represents a comprehensive approach by retaining spatial location information lost during pooling through a weighting mechanism.

Although numerous road extraction methods exist, the tabular review in Table 1 focuses on encoder–decoder architectures with structural similarity to the proposed model. These models integrate backbone feature extractors with context modules or multiscale fusion blocks, making them directly comparable to the proposed model.

Table 1. *Comparison of road extraction models in literature.*

Method	Performance	Strength	Reported limitation
D-LinkNet	Validation IoU (DeepGlobe): 0.6412	Efficient encoder-decoder backbone, less road discontinuity.	Cross-dataset robustness remains unclear due to evaluation only on a single benchmark. Limited multiscale feature fusion across encoder-decoder stages.
MSPFE-Net	Test IoU (Massachusetts Roads): 0.5909	False positive is reduced and the boundary is cleaner.	Generalization to different imaging conditions and geographical regions remains insufficiently explored. Feature enhancement module has no mechanism to adapt feature extraction to varying road widths.
MACU-Net	Test IoU: Massachusetts Roads 0.6457, DeepGlobe .6251	Good performance in complex scenes, multiscale feature aggregation gives strong representation.	No true global long-range contextual modelling. Only uses standard cross-entropy loss, making the model susceptible to imbalance road/background settings.
MECA-Net	Test IoU: Massachusetts Roads 0.6582, DeepGlobe 0.6515	Introduction of spatial and channel long-range contexts makes robust segmentation possible.	Minimal IoU improvement against the baseline model (less than 1% over U-Net in two datasets), signalling weak discriminative power for fine road structures.
RCFSNet	Test IoU: Massachusetts Roads 0.6424, DeepGlobe 0.6606	Competitive result on standard benchmark, good incremental ablation performance.	Too many custom modules increase model complexity. Does not explicitly address foreground-background imbalance.

1.2 Research gap and contributions

Despite substantial progress in deep learning-based road extraction, several challenges remain unresolved. Roads are thin and elongated structures that require accurate boundary delineation and continuity preservation. However, high-resolution remote sensing imagery frequently contains occlusions caused by vegetation, vehicles and buildings, along with significant appearance variations arising from surface materials, shadows and illumination conditions. In dense urban environments, roads often exhibit low contrast with surrounding paved areas, whereas in rural regions, they are partially obscured by vegetation. Although existing architectures employ multiscale feature extraction and attention mechanisms, many still struggle to simultaneously preserve fine-scale structural details and capture sufficient global context, resulting in fragmented road predictions and connectivity errors.

One potential approach to address these difficulties is to use a stronger feature extractor that can learn discriminative representations for distinguishing road and non-road features. ConvNeXt (Z. Liu et al., 2022) is one of this decade's advances in convolutional neural architectures and has demonstrated strong representational feature learning capabilities through its carefully crafted building blocks, which incorporate modern normalization layers and non-linear activation mechanisms. These characteristics make ConvNeXt a promising candidate for road extraction tasks, where accurate discrimination between roads and visually similar background objects requires both local detail preservation and contextual understanding.

Another recurring limitation identified in the reviewed literature is the lack of explicit treatment for road-pixel-to-background imbalance. For instance, in the Massachusetts Roads dataset, road pixels constitute only approximately 4.8% of the total pixel count, representing a foreground-background ratio of roughly 1:20. Evidence of the impact of this imbalance is provided by MSPFE-Net; replacing cross-entropy loss with Dice loss, which partially mitigates imbalance through overlap-based optimization, yielded a 2.19 percentage point improvement in IoU. This result indicates that class imbalance has a measurable and non-trivial effect on road extraction performance. Furthermore, D-LinkNet, MACU-Net and RCFSNet do not incorporate explicit class-frequency-aware weighting mechanisms during training, which may increase susceptibility to background-dominated optimization and reduce sensitivity to minority road pixels. These observations suggest that, despite advances in network architecture and feature representation, explicit imbalance-aware optimization remains underexplored in recent road extraction methods.

RCFSNet (Yang et al., 2023), a representative recent approach, improves robustness to occlusion through multi-level feature aggregation; however, it can still encounter difficulties in complex scenes due to the relatively limited representational capability of its encoder. Moreover, RCFSNet does not explicitly address foreground-background imbalance, leaving the learning process susceptible to background dominance. These observations indicate that existing methods still lack an integrated solution that simultaneously enhances feature representation, explicitly mitigates class imbalance and preserves structural continuity across multiple spatial scales.

To better understand these challenges, our study explores the following research questions:

- **RQ1**: Can a ConvNeXt-based encoder provide more discriminative feature representations for road extraction than the models used in existing encoder-decoder frameworks?
- **RQ2**: How does inverse-frequency-weighted optimization handle foreground-background imbalance to help the model identify minority road pixels?
- **RQ3**: Does a scale-sensitive decoder improve the extraction of roads that vary in their width and appearance?
- **RQ4**: Do these improvements, which function as orthogonal enhancements at the representation, optimization and decoding levels, collectively provide benefits that boost overall road extraction performance?

This work proposes **RoadNeXt-CSS (Context, Stages and Scale)**, a road extraction framework in which each component is designed to target specific challenges identified above. The main contributions are as follows:

(1) **ConvNeXt backbone adaptation:** We adapt ConvNeXt as the backbone encoder within an RCFSNet-derived road extraction framework. Unlike a straightforward backbone replacement, this integration requires architectural modifications to align the hierarchical feature representations of ConvNeXt with the decoder and FSFF-based feature fusion pipeline. The resulting encoder improves the extraction of discriminative spatial-semantic features from complex remote sensing scenes while preserving the road continuity modelling capabilities inherited from RCFSNet.

(2) **Scale-sensitive (ScaleSen) decoder refinement:** We incorporate a ScaleSen module into the decoder to improve the aggregation of multiscale road features. By adaptively weighting decoder feature maps according to their scale relevance, the proposed decoder enhances the representation of roads with varying widths and appearance characteristics while suppressing background interference.

(3) **Inverse frequency weighting (IFW) training strategy:** We introduce an inverse-frequency-weighted (IFW) loss strategy to address the severe foreground-background imbalance commonly encountered in road extraction datasets. By calibrating class contributions during optimization, the proposed strategy improves the learning of minority road pixels and reduces prediction bias towards the dominant background class.

These three contributions address orthogonal limitations of existing road extraction frameworks: encoder weakness (ConvNeXt), foreground-background class imbalance (IFW) and insufficient multiscale feature integration (ScaleSen). Since these components operate at the representation, optimization and decoding levels, respectively, their effects are complementary rather than redundant. Rather than introducing an entirely new architecture, RoadNeXt-CSS extends the RCFSNet framework through coordinated improvements across these three levels. By improving the accuracy and robustness of automated road extraction, RoadNeXt-CSS contributes to the

development of more reliable spatial information systems that support transportation planning, navigation, disaster response and related geospatial applications.

To facilitate transparency, reproducibility and future research, the implementation of RoadNeXt-CSS, including training and evaluation procedures, is publicly available at https://github.com/WaOnEmperoR/roadnext_css.

2 METHODOLOGY

2.1 ConvNeXt

ConvNeXt is a deep learning architecture that improves the performance of ResNet using techniques from the Vision Transformer (ViT). Developed by Facebook, it achieves 82% accuracy on ImageNet-1K, outperforming the Swin Transformer. Based on this superior performance, we adopted ConvNeXt as the backbone encoder for RoadNeXt-CSS. We hypothesize that using this state-of-the-art encoder provides better feature extraction capabilities than vanilla or traditional encoders. To corroborate this choice, we also refer to previous studies that have implemented ConvNeXt in remote sensing applications, namely AAUConvNeXt (P. Zhang et al., 2024) and ConvNeXt-ECANet (Perdana et al., 2024).

ConvNeXt improves ResNet through two main strategies: macro design optimization and micro design optimization. Macro design optimization adopts ViT training techniques, such as MixUp, CutMix and RandAugment, employs AdamW as the optimizer with additional regularization and modifies the convolutional structure to improve efficiency while increasing the number of channels. Micro design strategies replace ReLU with GeLU for more effective activation, reduce the number of activation functions, adjust normalization strategies and apply downsampling with layer normalization at each resolution change.

The core building block of ConvNeXt is the ConvNeXt block, which consists of a depthwise convolution with a 7×7 kernel followed by two 1×1 convolutions. A group of ConvNeXt blocks operating at the same spatial resolution before downsampling is referred to as a stage. A complete ConvNeXt model comprises one stem layer followed by four stages. ConvNeXt is available in four variants—Tiny, Small, Base and Large—each differing in the number of channels and the number of ConvNeXt blocks within each stage.

2.2 Encoder-decoder architecture

Road extraction can be described as a semantic segmentation task and is commonly addressed using encoder-decoder architecture. The encoder progressively transforms the input image into lower-dimensional feature representations to capture high-level semantic information, while the decoder gradually reconstructs these features to generate a pixel-wise segmentation map. In this study, the proposed RoadNeXt-CSS framework is built on RCFSNet, which is derived from the U-Net architecture (Ronneberger et al., 2015). U-Net is one of the most widely adopted encoder-decoder models due to its ability to achieve high segmentation accuracy even with limited training data. Although originally developed for biomedical image segmentation, U-Net has since been successfully applied to various semantic segmentation tasks, including remote sensing image analysis and road extraction.

As a bridge module, RoadNeXt-CSS adopts the multiscale context extraction (MSCE) module from RCFSNet to enhance contextual feature representation. Furthermore, the full-stage feature fusion (FSFF) module is retained in each skip connection to facilitate the integration of multilevel encoder features and preserve road continuity information. These modules are inherited from RCFSNet because of their demonstrated effectiveness in modelling contextual information and structural continuity in road extraction tasks.

In RoadNeXt-CSS, a fine-tuned ConvNeXt is employed as the encoder, replacing the ResNet34 backbone used in RCFSNet. The integration of ConvNeXt into the RCFSNet framework is nontrivial because ConvNeXt uses a patch-based stem followed by four hierarchical stages, whereas RCFSNet was originally designed around ResNet34 architecture. Consequently, direct reuse of the original skip connection design would lead to spatial resolution inconsistencies between encoder and decoder features. To address this issue, a transposed convolution layer is introduced to bridge the spatial resolution discrepancy and restore compatibility between encoder and decoder feature maps. In addition, the skip connection topology is modified by excluding the stem output from direct feature propagation. These architectural adaptations ensure that ConvNeXt feature maps are properly aligned with the FSFF

modules and decoder stages, enabling effective multilevel feature aggregation while preserving the structural continuity modelling capabilities inherited from RCFSNet.

In the decoder, a scale-sensitive neural network (ScaleSen) module (Tan et al., 2021) is incorporated to adaptively fuse features from multiple decoder levels and suppress less informative responses. This mechanism improves the use of multiscale information and enhances the extraction of road structures with varying widths and appearance characteristics. The details of the MSCE, FSFF, ConvNeXt integration and ScaleSen modules are presented in the following sections.

2.3 Full-stage feature fusion (FSFF)

The feature maps from the stem block and four stages of the ConvNeXt encoder (labelled E1 to E5) become input for the feature merging stage (FSFF) in the skip connection. The skip connections from E2, E3 and E4 contain three instances of FSFF in this architecture. At every stage, feature maps are used as input for each FSFF, as shown in **Figure 1**. The basic idea of FSFF is that the decoder receives feature maps at different stages, with sufficient hierarchical characteristics of the road so that the model can anticipate clearer boundaries.

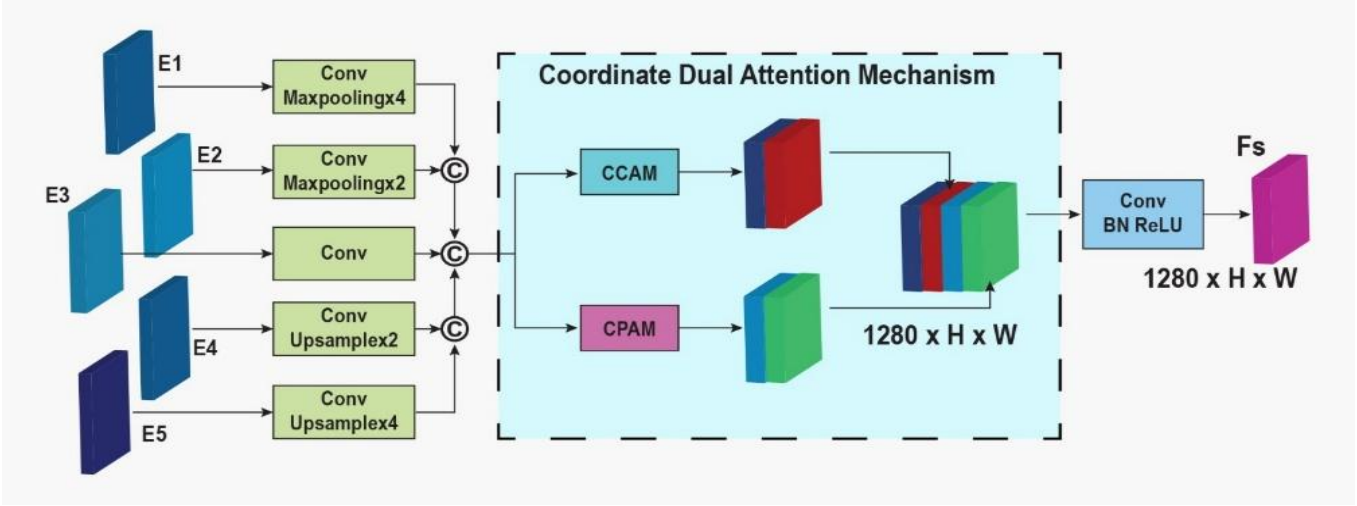


Figure 1. *Full-stage feature fusion (FSFF) module.*

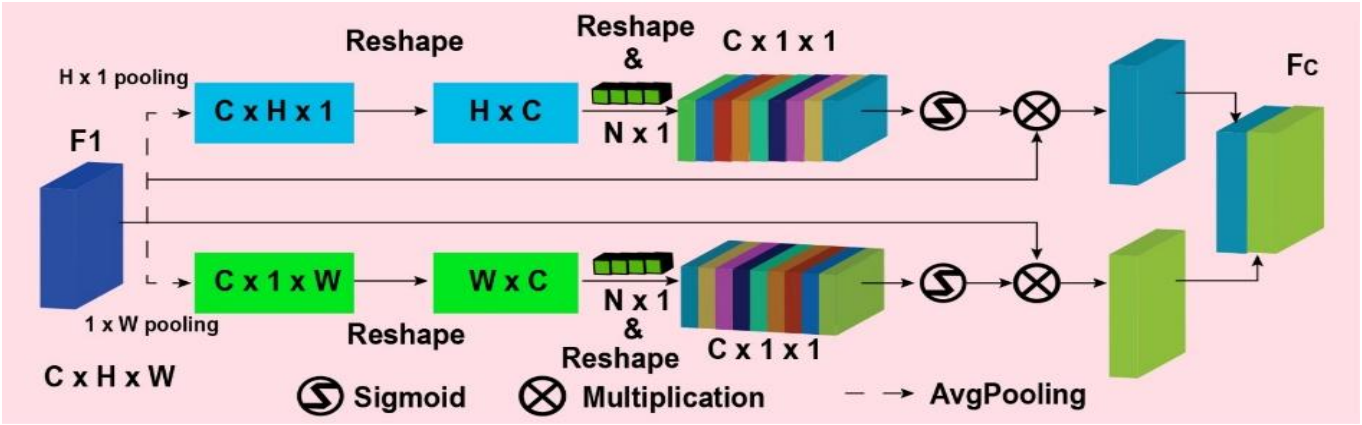


Figure 2. *Coordinate channel attention module (CCAM).*

In Figure 1, the concatenated features from each ConvNeXt stage are passed into the coordinate dual attention module (CDAM), which consists of two submodules: the coordinate channel attention module (CCAM) in Figure 2 and the coordinate positional attention module (CPAM) in Figure 3. CCAM, borrowing the implementation of ECA-Net (Q. Wang et al., 2020), aims to capture the cross-channel interaction in weight and height orientation. The global average and global maximum pooling for this feature is performed by CPAM, which uses the implementation of CBAM (Woo et al., 2018). The feature map of this operation is then passed into two convolutions in parallel with the kernels of size $(H, 1)$ to extract features in the horizontal directions and $(1, W)$ to extract features in the vertical

directions. All original feature channels are weighed because of this process. The channels of both CCAM and CPAM feature map results are merged by concatenating and passing them into one convolution block. This is the conclusion of an FSFF instance.

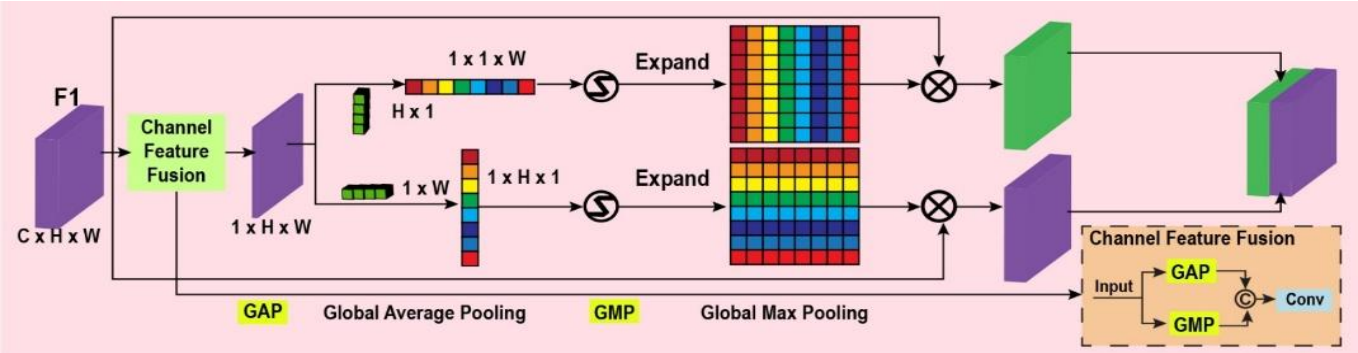


Figure 3. *Coordinate positional attention module (CPAM).*

To provide clarity, we explain all the modules mathematically. Let E_i denote the feature map extracted from the encoder stage i, where $i \in \{1, \ldots, 5\}$. To enable feature aggregation across multiple encoder stages, all the feature maps are first resized to a common spatial resolution (resolution of E_3):

$$\hat{E}_i = \mathcal{R}(E_i) \tag{1}$$

where $\mathcal{R}(\cdot)$ denotes upsampling (E_4, E_5) or downsampling (E_1, E_2).

The resized feature maps are then concatenated:

$$F1 = \text{Concat}(\hat{E}_1, \hat{E}_2, \hat{E}_3, \hat{E}_4, \hat{E}_5) \tag{2}$$

The coordinate channel attention module (CCAM) enhances channelwise feature representation by aggregating contextual information along the height and width dimensions separately. Given an input feature map ($F1$), average pooling is first performed along the height and width dimensions to generate two coordinate-aware descriptors. These descriptors are transformed and fused to produce two channel attention maps $F1_h$ and $F1_w$, which are then concatenated to produce the final feature map. This procedure can be formulated as:

$$F1_h = F1 \otimes \sigma\big(P_h(F1)\big) \tag{3}$$

$$F1_w = F1 \otimes \sigma\big(P_w(F1)\big) \tag{4}$$

$$F1_{\text{CCAM}} = \text{Concat}(F1_h, F1_w) \tag{5}$$

where $P_w(\cdot)$ and $P_h(\cdot)$ denote widthwise and heightwise average pooling operations, respectively, $\sigma(\cdot)$ denotes the sigmoid activation function and $\otimes$ denotes elementwise multiplication.

The coordinate positional attention module (CPAM) is designed to capture positional dependencies along both horizontal and vertical directions. First, global average pooling (GAP) and global max pooling (GMP) are applied in parallel to the input feature map ($F1$) and their outputs are concatenated to form the coordinate feature fusion representation ($F1_{\text{CFF}}$), as shown in Equation (6). The fused feature map is then processed by two parallel convolution branches. A horizontal convolution with kernel size $(H \times 1)$ generates the horizontal attention descriptor (Y_h) (Equation (7)), whereas a vertical convolution with kernel size $(1 \times W)$ generates the vertical attention descriptor (Y_w) (Equation (8)). The resulting descriptors are activated using the sigmoid function and subsequently expanded through broadcasting operations to match the spatial dimensions of the input feature map. Finally, the expanded horizontal and vertical attention features are concatenated to produce the CPAM output ($F1_{\text{CPAM}}$), as defined in Equation (9). Through this mechanism, CPAM enhances the ability of the model to capture long-range positional information while preserving spatial continuity along both coordinate directions.

$$F1_{\mathrm{CFF}} = \mathrm{Concat}(\mathrm{GAP}(F1), \mathrm{GMP}(F1)) \quad (6)$$

$$Y_h = \sigma\left(\mathrm{Conv}_{H\times 1}(F1_{\mathrm{CFF}})\right) \quad (7)$$

$$Y_w = \sigma\left(\mathrm{Conv}_{1\times W}(F1_{\mathrm{CFF}})\right) \quad (8)$$

$$F1_{\mathrm{CPAM}} = \mathrm{Concat}(\mathrm{Expand}_H(Y_h), \mathrm{Expand}_W(Y_w)) \quad (9)$$

Then, these two feature maps are concatenated and projected through a convolution layer to form the final output of the FSFF module:

$$F_{\mathrm{CDAM}} = \mathrm{Conv}\left(\mathrm{Concat}(F1_{\mathrm{CCAM}}, F1_{\mathrm{CPAM}})\right) \quad (10)$$

2.4 Multiscale context extraction module (MSCE)

Placed on the bridge, the MSCE in Figure 4 is used to capture the relationship between the body of the road and the surrounding area so that it can obtain characteristics, improve road body connectivity and reduce disconnection, for example, due to occlusion caused by trees. Dilated convolution is used by MSCE to increase its receptive field, enabling it to cover more information. In the deepest layer, the feature map will be copied 5 times: 3 for the dilated convolution block branch and 2 for the pooling operation.

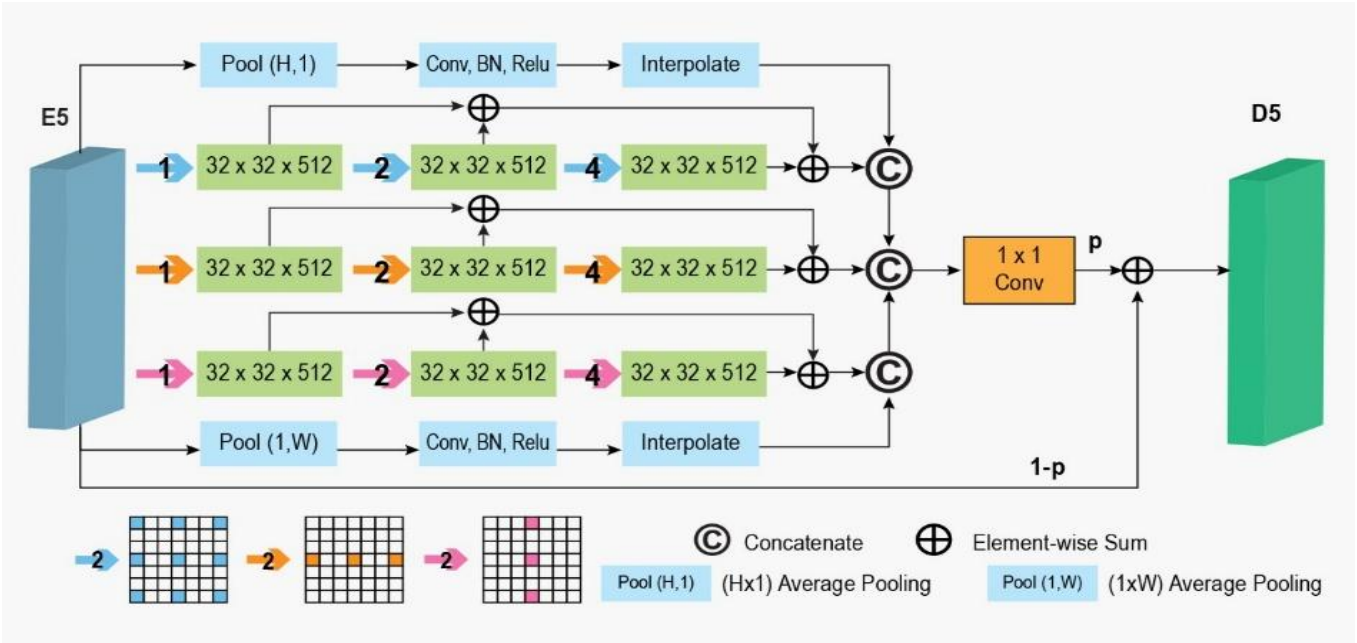


Figure 4. *Multiscale context extraction (MSCE) module.*

2.5 Decoder with scale-sensitive module

The decoder at the level n receives input from the FSFF at the skip connection level n and the decoder level $(n-1)$. Upsampling by scale two is performed on the feature map of the decoder $(n-1)$ to make it the same size as the feature map from the FSFF. Next, both features are concatenated and integrated with convolution, batch normalization (BN) and rectified linear unit (ReLU) activation functions. This process produces the final feature map for the n-th level decoder D_n.

In RoadNeXt-CSS, we integrated the scale-sensitive neural network to fuse the feature maps from each level of the decoder. The scale-sensitive neural network performs weighting at each level of the decoder using a weight tensor. In this model, weighting is performed on feature D_4, D_3, D_2 and D_1 with trainable tensor parameters w_4, w_3, w_2 and w_1 respectively. This weighting is used to make noisy features less specific and to give greater importance to features that play a greater role in recognizing road networks.

For the feature map D_5, no weighting mechanism is performed because D_5 represents the global characteristics of the road network in the image. Grad-CAM visualization showed that D_5 consistently produced strong activations over the overall road structure, indicating that the feature map already captured sufficient global semantic information. This observation was further supported by preliminary experiments, where introducing a trainable weighting term for D_5 was evaluated but consistently led to a slight performance degradation in both datasets. We hypothesize that, due to its lowest spatial resolution among the decoder features, increasing the contribution of D_5 reduces the relative influence of finer-grained spatial details extracted from higher-resolution feature maps.

The final segmentation result is obtained using the sigmoid function in Equation (11).

$$\text{Predict} = \sigma\left(D_5 + \sum_{i=1}^{4} w_i \cdot D_i\right) \tag{11}$$

We show the general diagram of the RoadNeXt-CSS architecture in Figure 5. This highlights the usage of the ConvNeXt encoder, MSCE, FSFF and scale-sensitive network in the decoder.

2.6 Weighted loss function

Many deep learning-based road extraction models, such as RCFSNet, use binary cross entropy (BCE) (Equation (12)) and or Dice coefficient (Equation (13)) (Sudre et al., 2017) as the loss function to train the model.

$$L_{CE} = -\frac{1}{N}\sum_{c=1}^{C}\sum_{i=1}^{N} g_{i,c} \cdot \log p_{i,c} \tag{12}$$

$$L_{Dice=}1 - \frac{2\sum_{c=1}^{C}\sum_{i=1}^{N} g_{i,c} \cdot p_{i,c}}{\sum_{c=1}^{C}\sum_{i=1}^{N} g_{i,c} + \sum_{c=1}^{C}\sum_{i=1}^{N} p_{i,c}} \tag{13}$$

$$L_{combination} = L_{CE} + L_{Dice} \tag{14}$$

where $g_{i,c}$ is the ground truth label at the i-th pixel for the class c, $p_{i,c}$ is the predicted value of the i-th pixel for the class c. N is the total number of pixels in the image. For the case of binary classification, the value is $C = 2$.

In this study, we adopted the inverse frequency weighting (IFW) (Sugino et al., 2021) in the loss calculation. IFW is one of the most widely used strategies for weighting classes based on their class frequencies. The working principle is that the smaller the size of the target object, the greater the weight value. IFW can help overcome the problem of data imbalance, making the model focus on large target objects and forget about small targets. The modified loss function with IFW is described in Equations (15), (16) and (17).

$$L_{CE}^{\text{inv}} = -\frac{1}{N}\sum_{c=1}^{C}\frac{\sum_{i=1}^{N} g_{i,c} \cdot \log p_{i,c}}{\sum_{i=1}^{N} g_{i,c}} \tag{15}$$

$$L_{Dice}^{\text{inv}} = 1 - \frac{2\sum_{c=1}^{C}\frac{\sum_{i=1}^{N} g_{i,c} \cdot p_{i,c}}{\sum_{i=1}^{N} g_{i,c}}}{\sum_{c=1}^{C}\frac{\sum_{i=1}^{N}(g_{i,c} + p_{i,c})}{\sum_{i=1}^{N} g_{i,c}}} \tag{16}$$

$$L_{combination}^{\text{inv}} = L_{CE}^{\text{inv}} + L_{Dice}^{\text{inv}} \tag{17}$$

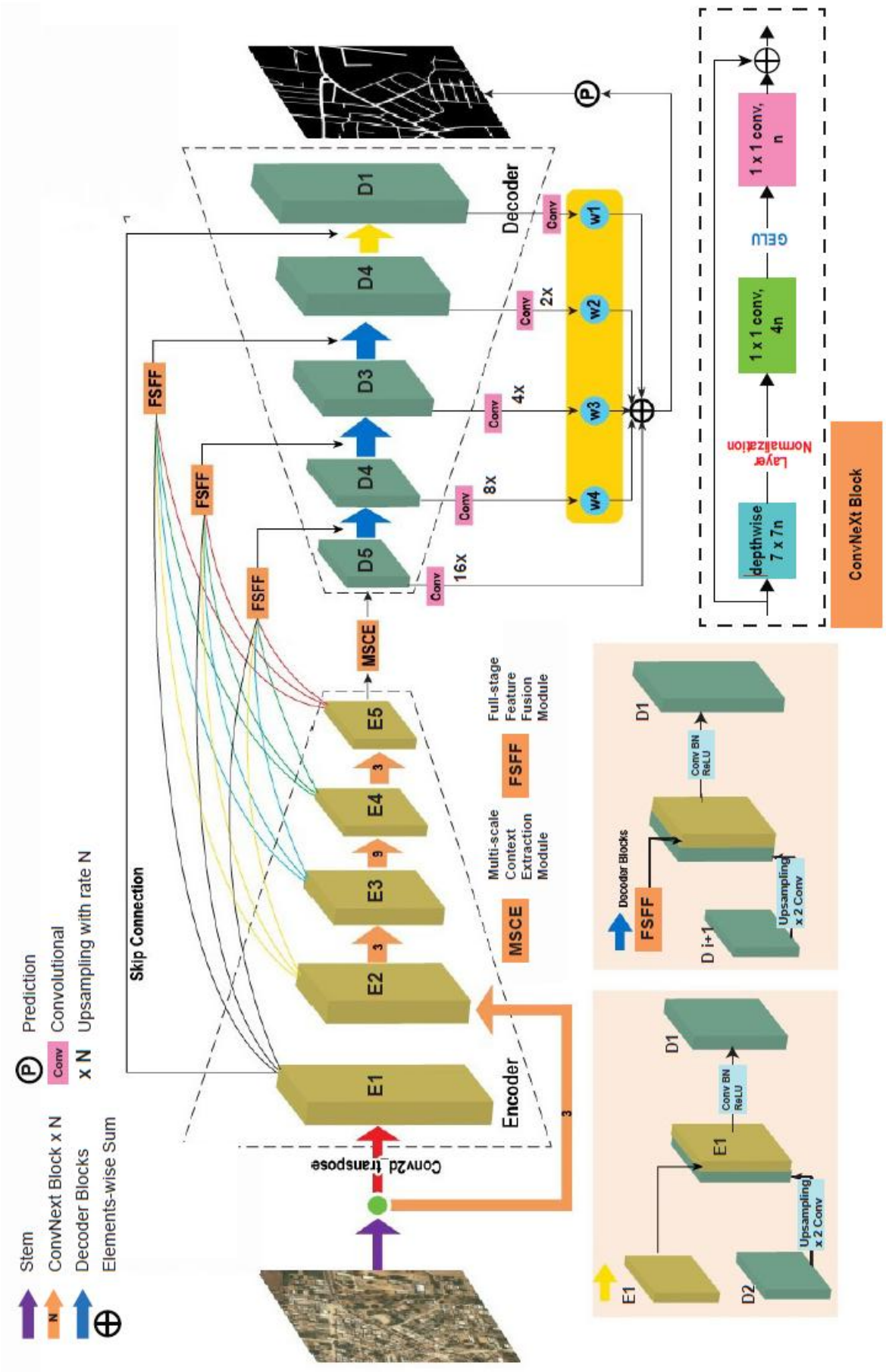


Figure 5. *Proposed RoadNeXt-CSS architecture.*

Since road extraction is formulated as a binary segmentation task, consisting of only two classes (road = 1 and background = 0), the general multiclass IFW formulation in Equations (15)-(17) can be simplified by omitting the class index (c). In this case, the weighted BCE and Dice losses are implemented using the positive-class weight parameter, denoted as (w_{pos}) (corresponding to PyTorch **pos_weight**). The value of (w_{pos}) is calculated as the ratio between the number of background pixels and the number of road pixels in the training dataset. This formulation can be directly implemented using the built-in loss functions of PyTorch. Accordingly, the binary weighted loss functions used in this study are expressed in Equations (18), (19) and (20).

$$L_{WBCE} = -\frac{1}{N}\sum_{i=1}^{N} w_{pos} \cdot g_i \cdot \log(p_i) + (1 - g_i) \cdot \log(1 - p_i) \tag{18}$$

$$L_{WDice} = 1 - \frac{2\left(w_{pos} \cdot \sum_i (g_i \cdot p_i) + \sum_i ((1-g_i) \cdot (1-p_i))\right)}{\left(w_{pos} \cdot \sum_i (g_i + p_i)\right) + \sum_i ((1-g_i) + (1-p_i))} \tag{19}$$

$$L_{combination} = L_{WBCE} + L_{WDice} \tag{20}$$

3 EXPERIMENT SETTING

3.1 Dataset

The model will be tested in two remote sensing datasets. The first is the Massachusetts Roads dataset (MRD). This dataset consists of 1171 aerial images of the state of Massachusetts, United States. Each image is 1500×1500 pixels in size and covers an area of 225 km^2. This dataset displays many regional variations, namely urban, suburban and rural areas with a total area of 2600 km^2. The dataset is divided into 1108 images for training, 14 for validation and 49 for testing. We resized it into 1024×1024 pixels.

The second dataset is DeepGlobe, taken from a competition at cVPR 2018. DeepGlobe images were acquired from DigitalGlobe satellites with a pixel resolution of 50 cm. The dataset consists of 6226 images at 1024×1024 pixels for training, 1243 for validation and 1101 for testing acquired in India, Thailand and Indonesia. This dataset is very challenging since it encompasses many countries and landscapes, including cities, villages, farmlands and forested areas. The labelled images are in greyscale format with black pixels as background and white pixels as road bodies. Small roads in agricultural areas are not annotated. Since the validation and testing images are not annotated, following the procedure described in the RCFSNet paper, 5500 images will be taken as training data and 726 images as test data. The splitting strategy is deterministic random sampling (random_state=2024) from images labelled as split='train' in the provided metadata file. The exact split used in this study will be released through the project GitHub repository to facilitate reproducibility.

Although the Massachusetts Roads and DeepGlobe datasets are not newly released, they remain two of the most widely used and publicly available benchmarks for road extraction from remote sensing images. Their rich annotations, diversity of geographical coverage and community acceptance make them highly suitable for evaluating and comparing model performance. Numerous recent studies (e.g., Y. Zhang et al., 2023; Akhtarmanesh et al., 2024; Ahmed et al., 2024; Z. Wei & Zhang, 2023) continue to utilize these datasets to test novel architectures and techniques. Using these, we ensure that our proposed method can be directly compared with existing work under consistent and recognized evaluation settings.

3.2 Evaluation metrics

This model will be evaluated using several metrics, namely IoU, precision, recall and F1 score. These metrics use four reference values: true positive (TP), true negative (TN), false positive (FP) and false negative (FN). The first metric is intersection over union (IoU). This metric calculates the ratio of how much the ground truth and predicted pixels overlap. IoU is defined as

$$IoU = \frac{\mathrm{TP}}{\mathrm{TP+FN+FP}} \tag{21}$$

The other metrics are precision (the accuracy of the model in predicting positive values), recall (the ability of the model to detect actual positive cases) and F1 score (the harmonic mean of precision and recall). The range of F1 score values is [0, 1], where the best value is 1. The reported metrics are the average values of the four metrics for the test images in the Massachusetts Roads and DeepGlobe datasets.

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \tag{22}$$

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{23}$$

$$F1 = \frac{2\times Precision\times Recall}{Precision+Recall} \tag{24}$$

3.3 Hyperparameter

The experiment was conducted in NVIDIA DGX a100 with 8 GPUs, in which only 2 GPUs can be used by users simultaneously. We used the Adam optimizer (learning rate = 2×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$) and the batch size was 6. The learning rate would be reduced by the factor of one-fifth after 3 consecutive epochs if the validation accuracy was not increased and the learning rate was still greater than 5×10^{-7}. Training would stop if there was no improvement in six consecutive epochs. Input images were resized to 1024 × 1024 and normalized via $\hat{x} = (x/255) \cdot 3.2 - 1.6$, like in the D-LinkNet normalization technique. Online data augmentation was applied during training, comprising random HSV jitter (hue ∈ [−30°, 30°], saturation ∈ [−5, 5], value ∈ [−15, 15]), random shift/scale/aspect warp (±0.1), random horizontal flip, random vertical flip and random 90° rotation, each applied independently with a probability of 0.5. We ran the training for 100 epochs on Massachusetts Roads and 50 epochs on DeepGlobe. Training completed in approximately 11.9 hours (100 epochs) on Massachusetts Roads and 14.8 hours (50 epochs) on DeepGlobe.

3.4 Ablation studies

In ablation studies, we test the effect of the type of encoder, loss function and use of scale-sensitive network in decoder on the performance of road extraction model. Here, we use two kinds of encoders from the ConvNeXt family, namely ConvNeXt-Tiny and ConvNeXt-Small (each having a number of blocks in each stage (3,3,9,3) and (3,3,27,3), respectively). We also tested the effect of the use of weighted (Dice + BCE) loss function in training compared with the original (Dice + BCE) loss function. Lastly, we test the effect on the use of scale-sensitive networks compared with the original architecture, where the feature fusion from each decoder layer is simply done by adding. In this experiment, we used RCFSNet as the baseline, since RoadNeXt-CSS can be considered an improvement of RCFSNet (RoadNeXt-CSS also uses components from RCFSNet, namely MSCE and FSFF). It uses ResNet34 as an encoder, trained with (Dice + BCE) loss and did not use a scale-sensitive network in the decoder. Since FSFF and MSCE are inherited from RCFSNet and are not modified in this work, they are treated as fixed components and are not included in the ablation study.

Table 2. *Scenario settings.*

Scenario	Base encoder	Loss function	Use scale-sensitive decoder?
0	Baseline (RCFSNet)		
1	ConvNeXt-Tiny	Dice+BCE	No
2	ConvNeXt-Tiny	Weighted (Dice+BCE)	No
3	ConvNeXt-Tiny	Dice+BCE	Yes
4	ConvNeXt-Tiny	Weighted (Dice+BCE)	Yes
5	ConvNeXt-Small	Dice+BCE	No
6	ConvNeXt-Small	Weighted (Dice+BCE)	No
7	ConvNeXt-Small	Dice+BCE	Yes
8	ConvNeXt-Small	Weighted (Dice+BCE)	Yes

For the sake of fair comparison, the same evaluation protocol was adopted throughout all the ablation experiments and comparisons with other methods. Specifically, test-time augmentation (TTA) was applied during inference, where predictions from the original image and its flipped variants were aggregated to obtain the final segmentation result. The combinations of these components are organized into eight scenarios, which are detailed in Table 2.

4 RESULTS

4.1 Quantitative results of experiments

In this section, we present our findings from this experiment. First, we ran the baseline model on our machine with the configuration described above to obtain baseline (Scenario 0) performances. The improvements in Scenarios 1 to 8 are projected to achieve better performance than the baseline. The results of our experiments on Massachusetts Roads are displayed in Table 3 and on DeepGlobe in Table 4.

In Scenario 1, replacing the ResNet34 encoder (as in RCFSNet) with a more powerful model, ConvNeXt-Tiny, improves the IoU on Massachusetts Roads by 0.0097 points and on DeepGlobe by 0.0257 points. This means that the ConvNeXt-Tiny encoder is better capable of extracting road features. It achieves an IoU of 0.6521 on Massachusetts Roads and 0.6863 on DeepGlobe.

Further improvement is observed in Scenario 2, where inverse frequency weighting (IFW) is incorporated into the loss function. As an initial estimate, the **pos_weight** parameter was derived from the foreground-to-background pixel ratio, since background pixels substantially outnumber road pixels in both the Massachusetts Roads and DeepGlobe datasets. However, our experiments show that the best segmentation performance is achieved with lower **pos_weight** values than those suggested by class frequencies alone. This indicates that the optimal weight is influenced not only by class imbalance but also by dataset-specific characteristics, such as scene complexity, visual similarity between road and non-road regions and the resulting precision–recall tradeoff. Moreover, the Dice loss already partially compensates for imbalance and roads are not uniformly distributed (they cluster spatially), making the simple ratio too aggressive. Therefore, while the class ratio provides a useful starting point for selecting **pos_weight**, the parameter should be regarded as a dataset-dependent hyperparameter that requires empirical tuning to achieve optimal performance.

Therefore, we evaluated several **pos_weight** values to investigate the effect of inverse frequency weighting on road extraction performance. For the Massachusetts Roads dataset, the tested values were 3.0, 5.0, 6.0, 10.0 and 20.0, whereas for DeepGlobe, they were 2.0, 2.5, 3.0 and 5.0. We observed that precision generally decreased as **pos_weight** increased, whereas recall tended to increase, indicating a trade-off between false positive and false negative predictions. The best overall performance, as measured by IoU and F1 score, was achieved at **pos_weight** = 5.0 for Massachusetts Roads and **pos_weight** = 2.5 for DeepGlobe. These results suggest that the optimal weighting coefficient is not determined solely by the foreground-to-background pixel ratio but is also influenced by dataset-specific characteristics and the resulting precision–recall balance. Compared with Scenario 1, the selected **pos_weight** values improved the IoU by 0.0203 on Massachusetts Roads and 0.0058 on DeepGlobe, achieving IoU scores of 0.6724 and 0.6921, respectively.

Scenario 3 integrates the scale-sensitive module in the decoder part of the model with the ConvNeXt-Tiny encoder. We successfully demonstrate that the integration of this scale-sensitive module improves the IoU and F1 scores, compared with the model with the ConvNeXt-Tiny encoder without the scale-sensitive module, as in Scenario 1. It achieves an IoU of 0.6542 on Massachusetts Roads and 0.6934 on DeepGlobe.

Scenario 4 integrates the scale-sensitive module in the decoder part with the ConvNeXt-Tiny encoder and trains it with the weighted loss function. We used the optimal **pos_weight** value obtained in Scenario 2 to train the model on the Massachusetts Roads and DeepGlobe datasets. This experiment further achieves a higher IoU and F1 score for the respective dataset. This indicates the capabilities of scale-sensitive modules in optimizing the capabilities of the decoder to enhance important features that are useful for road segmentation or extraction. It achieves IoU of 0.6736 on Massachusetts Roads and 0.6965 on DeepGlobe.

Table 3. *Recapitulation of performances on Massachusetts Roads.*

Scenario	Components	IoU	Precision	Recall	F1 score
0	RCFSNet	0.6410	0.8709	0.7082	0.7781
1	ConvNeXt-Tiny	0.6521	0.8767	0.7186	0.7868
2	ConvNeXt-Tiny+IFW	0.6724	0.8141	0.7932	0.7934
3	ConvNeXt-Tiny+ScaleSen	0.6542	0.8735	0.7228	0.7880
4	ConvNeXt-Tiny+IFW+ScaleSen	0.6736	0.8215	0.7885	0.8028
5	ConvNeXt-Small	0.6558	**0.8850**	0.7163	0.7898
6	ConvNeXt-Small+IFW	0.6746	0.8083	0.8039	0.8037
7	ConvNeXt-Small+ScaleSen	0.6595	0.8782	0.7258	0.7924
8	ConvNeXt-Small+IFW+ScaleSen	**0.6779**	0.8105	**0.8047**	**0.8061**

Table 4 *Recapitulation of performances on DeepGlobe.*

Scenario	Components	IoU	Precision	Recall	F1 score
0	RCFSNet	0.6606	0.7855	0.8097	0.7859
1	ConvNeXt-Tiny	0.6863	0.7887	0.8409	0.8055
2	ConvNeXt-Tiny+IFW	0.6921	0.7821	0.8595	0.8102
3	ConvNeXt-Tiny+ScaleSen	0.6934	0.8034	0.8361	0.8105
4	ConvNeXt-Tiny+IFW+ScaleSen	0.6965	0.7950	0.8510	0.8130
5	ConvNeXt-Small	0.6895	**0.8063**	0.8260	0.8067
6	ConvNeXt-Small+IFW	0.6974	0.7899	0.8553	0.8121
7	ConvNeXt-Small+ScaleSen	0.6921	0.8020	0.8360	0.8087
8	ConvNeXt-Small+IFW+ScaleSen	**0.6984**	0.7871	**0.8629**	**0.8146**

Scenario 5 replaces ResNet34 with ConvNeXt-Small as the base encoder in RCFSNet. From this experiment, the performance of RCFSNet with ConvNeXt-Small encoder can outperform the baseline model and model with ConvNeXt-Tiny encoder (Scenario 1) in both test datasets. This means that the ConvNeXt-Small encoder is better capable of extracting road features. Compared with the baseline model, the model in Scenario 5 yields IoU higher by 0.0148 points on Massachusetts Roads and 0.0289 points on DeepGlobe. It achieves an IoU of 0.6558 on Massachusetts Roads and 0.6895 on DeepGlobe.

Scenario 6 uses RCFSNet with ConvNeXt-Small encoder and modifies the loss function with a weighting mechanism, similar to the experiment in Scenario 2. As our test using **pos_weight** of 5.0 on Massachusetts Roads did not yield higher IoU than the experiment results in Scenario 3, we tried several other **pos_weight** values (6.0, 7.5 and 10.0). In this scenario, we obtained the best IoU and F1 score at the **pos_weight** value of 6.0. A similar case occurred with DeepGlobe, where we tried several values (2.5, 3.0, 4.0 and 5.0). The best IoU and F1 score were obtained using a **pos_weight** value of 3.0.

Scenario 7 uses similar settings as Scenario 3, but we use ConvNeXt-Small as the encoder instead of ConvNeXt-Tiny. The combination of ConvNeXt-Small as base encoder and scale-sensitive module is proven to increase the IoU and F1 score in both datasets, compared to RCFSNet with ConvNeXt-Small encoder without the scale-sensitive module in Scenario 5.

The setting for Scenario 8 is similar to that for Scenario 4 with the integration of the scale-sensitive decoder; except we are using ConvNeXt-Small instead of ConvNeXt-Tiny as the base encoder. We use the optimal **pos_weight** value obtained in Scenario 6 to train the model on the Massachusetts Roads and DeepGlobe datasets. As expected, this experiment yields a higher IoU and F1 score than the previous scenario, which once again demonstrates that the scale-sensitive module could enhance the capability of the base encoder in the road extraction task.

4.2 Analysis

Based on the experimental results of Scenarios 1 to 8, the use of a more robust encoder, a modified loss function (IFW) and the integration of a scale-sensitive module in the decoder have been proven to gradually improve the performance of the RoadNeXt-CSS model. It can be concluded that ConvNeXt-Small is a better encoder than ConvNeXt-Tiny and ResNet34. We also see an increase in precision and recall, showing the capabilities of the ConvNeXt family encoder to extract better features than the baseline model, which uses ResNet34.

The use of a weighted loss function also improves the performance of road extraction, as it reduces the false negative rate. The IoU and F1 score values increased compared with the model trained without weighted loss function. We also see that the increase in recall means that the model can now focus more on pixels of the foreground class. However, the optimal weight is dataset-dependent and does not necessarily follow the foreground-to-background ratio. This indicates that while IFW is effective, its performance is sensitive to weight selection and thus requires empirical tuning.

The scale-sensitive decoder successfully improves RCFSNet performance, as evidenced by the increase in the IoU and F1 score values in both models trained with the loss functions (Dice + BCE) and weighted (Dice + BCE) in both datasets. Thus, it is proven that the use of a scale-sensitive decoder can make the model more focused on important features from various decoder levels, as well as suppress features that contain noise.

The integration of the model with a scale-sensitive decoder and weighted (Dice + BCE) loss further strengthens the ability of the model to extract roads. The precision value increases again and reduces false positives due to the use of a scale-sensitive decoder, so when combined with a high recall rate, it will achieve the optimal IoU value. We observed this pattern in both the Massachusetts Roads and DeepGlobe datasets (using the ConvNeXt-Tiny encoder), and the highest IoU was achieved when the model used the ConvNeXt-Small encoder. It achieved an IoU of 0.6779 on Massachusetts Roads and 0.6984 on DeepGlobe. This result far outperforms the baseline, which only yields an IoU of 0.6424 and 0.6606 on Massachusetts Roads and DeepGlobe, respectively. This confirms the advantage of the model in Scenario 8 and we designated it as the final **RoadNeXt-CSS** model.

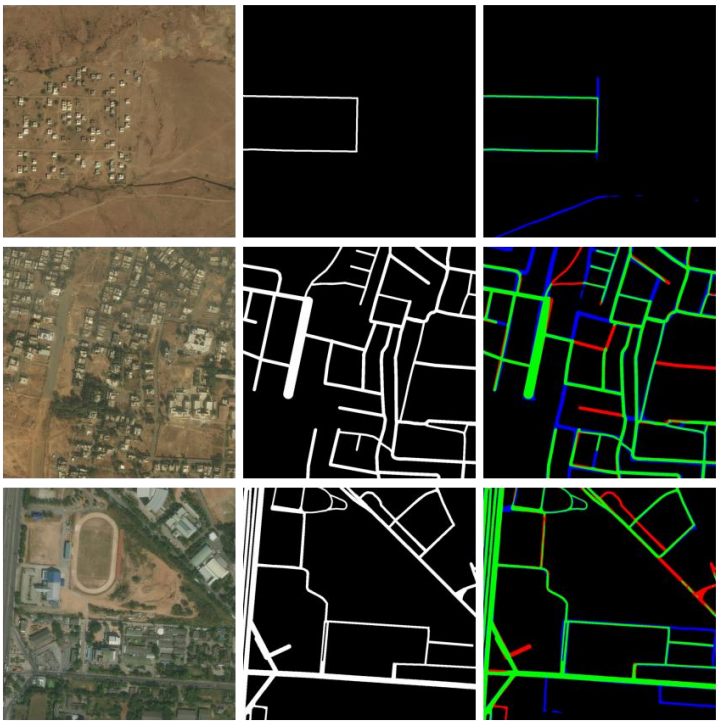


Figure 6. *Example of false positive predictions.*

To further investigate the limitations of RoadNeXt-CSS, we conducted a qualitative error analysis on representative test samples, as shown in Figure 6. Green pixels denote true positives (TP), red pixels denote false negatives (FN) and blue pixels denote false positives (FP). Most false positives occur on elongated structures with road-like

characteristics, such as drainage channels, artificial boundaries and visually plausible road segments that may be absent from the reference annotations.

Another observed failure mode is the overestimation of road continuity in occluded regions. In several cases, the model predicts road segments beneath dense tree canopies despite the road surface not being directly visible. This suggests that the model has learned a strong connectivity prior that favours continuous road networks. Overall, the false positives are generally structured and semantically meaningful rather than isolated noise, indicating that the model has learned robust road-related spatial priors while still facing challenges in distinguishing roads from visually similar linear objects.

4.3 Significance test (paired *t*-test)

Tables 5 and 6 present the paired *t*-test results for the Massachusetts Roads and DeepGlobe datasets, respectively, using a matrix representation based on scenario labels (S0–S8). Each matrix cell reports the *p*-value obtained from a paired comparison of IoU scores between two scenarios, where the scenario listed in the row is compared against the scenario listed in the column. A statistically significant performance difference is identified when the *p*-value satisfies the predefined significance criterion ($p < 0.05$), which is highlighted in bold. This representation enables a systematic evaluation of the contribution of individual design choices: encoder upgrades, inverse-frequency-weighted loss and scale-sensitive decoding across all scenario combinations and across both datasets.

The paired *t*-test was conducted using per-image IoU values from the test dataset. For each test image, the IoU obtained from the baseline model was paired with the IoU obtained from the comparison model. Thus, the sample size corresponded to the number of test images in the dataset. The test evaluated whether the mean difference in IoU between the two models was significantly different from zero.

The paired *t*-test results for the Massachusetts Roads dataset indicate that the inverse-frequency-weighted (IFW) loss function has a substantial influence on model performance. Statistically significant differences are consistently observed between models trained without and with IFW, including comparisons between (S1, S2) for the ConvNeXt-Tiny encoder and (S5, S6) for the ConvNeXt-Small encoder. In addition, the performance gains associated with IFW remain statistically significant when the scale-sensitive decoding module is incorporated, as shown in (S3, S4) and (S7, S8). These results suggest that loss reweighting addresses the pronounced class imbalance in the Massachusetts Roads dataset.

Table 5. *p-values for experiments on Massachusetts Roads.*

Methods	S0	S1	S2	S3	S4	S5	S6	S7	S8
S0		**1.78e-02**	**<0.001**	**8.0e-05**	**<0.001**	**1.65e-05**	**<0.001**	**1.00e-05**	**<0.001**
S1			**5.0e-05**	5.99e-01	**2.0e-05**	3.58e-01	**4.00e-05**	**3.00e-02**	**<0.001**
S2				**1.0e-05**	6.53e-01	**6.0e-05**	6.10e-01	**1.10e-03**	1.69e-01
S3					**<0.001**	6.43e-01	**3.0e-05**	7.31e-02	**<0.001**
S4						**1.0e-05**	7.87e-01	**7.0e-05**	2.52e-01
S5							**<0.001**	**3.29e-02**	**<0.001**
S6								**3.0e-05**	2.01e-01
S7									**<0.001**
S8									

Improvements were also observed when upgrading the encoder from ResNet34 to ConvNeXt-based backbones. The transitions from ResNet34 to ConvNeXt-Tiny (S0, S1) and to ConvNeXt-Small (S0, S5) yield statistically significant performance gains, confirming the benefit of enhanced feature representation. However, the comparison between ConvNeXt-Tiny and ConvNeXt-Small (S1, S5) does not show a statistically significant difference, indicating that increasing encoder capacity alone provides limited additional benefit for this dataset.

The effect of the scale-sensitive decoding module is less consistent. When applied without IFW, the module does not lead to significant improvement for ConvNeXt-Tiny (S1, S3), whereas a significant gain is observed for ConvNeXt-

Small (S5, S7). These findings indicate that multiscale decoding alone is insufficient to overcome the dominant class imbalance issue. Nevertheless, the combined use of encoder scaling, IFW and scale-sensitive decoding does not reduce performance. Instead, these components contribute complementary effects that improve model robustness, even when the resulting gains are modest.

Table 6. *p-values for experiments on DeepGlobe.*

Methods	S0	S1	S2	S3	S4	S5	S6	S7	S8
S0		**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
S1			**2.57e-02**	**4.7e-04**	**1.0e-05**	8.31e-01	**2.00e-05**	**2.11e-02**	**<0.001**
S2				6.00e-01	6.79e-02	7.56e-02	5.31e-02	9.60e-01	**7.91e-03**
S3					1.82e-01	**7.07e-03**	1.12e-01	5.72e-01	**2.14e-02**
S4						**3.80e-04**	7.23e-01	9.73e-02	3.78e-01
S5							**7.00e-05**	**4.04e-02**	**3.00e-05**
S6								**4.27e-02**	7.11e-01
S7									**2.06e-02**
S8									

The paired *t*-test results for the DeepGlobe dataset reveal a different performance pattern compared with the Massachusetts Roads dataset. In this case, the scale-sensitive decoding module shows a stronger and more consistent impact on performance. Statistically significant improvements were observed when introducing the scale-sensitive module without IFW for both ConvNeXt-Tiny (S1, S3) and ConvNeXt-Small (S5, S7) encoders. This indicates that explicit multiscale feature aggregation is important for modelling the wide variation in road widths and spatial structures in the DeepGlobe imagery.

The IFW loss function also contributes to performance improvements, as evidenced by significant differences in (S1, S2) and (S5, S6). However, its effect becomes less pronounced when scale-sensitive decoding is already applied. In particular, the comparison between (S3, S4) using ConvNeXt-Tiny does not yield a statistically significant improvement, suggesting that the benefits of loss reweighting may saturate once multi-scale features are effectively captured.

Encoder upgrades from ResNet34 to ConvNeXt-Tiny (S0, S1) and ConvNeXt-Small (S0, S5) result in statistically significant gains, confirming the general effectiveness of modern encoder architectures. Similar to the Massachusetts Roads dataset, further scaling from ConvNeXt-Tiny to ConvNeXt-Small (S1, S5) does not produce a significant performance difference. While some individual architectural modifications yield only modest improvements when evaluated in isolation, the ablation study demonstrates that each component contributes complementary benefits. Overall, the results indicate that encoder scaling, IFW and scale-sensitive decoding do not counteract each other. Instead, their combined use yields complementary improvements that enhance model stability and generalization, although the magnitude of performance gains varies depending on the dataset characteristics.

4.4 Computational complexity

The computational complexity comparison of the RCFSNet baseline and RoadNeXt-CSS is displayed in Table 7.

Table 7. *Computational complexity of RCFSNet and RoadNeXt-CSS.*

Model	Params (M)	Inference time (ms)	FLOPs (G)	FPS	Memory (GB)
RCFSNet	76.74	729.10	37.67	26.55	1.590
RoadNeXt-CSS	174.21	1637.58	85.79	11.66	2.649

As shown in Table 7, RoadNeXt-CSS exhibits higher computational complexity than the baseline RCFSNet. Specifically, the number of parameters increased by 127%, while memory consumption increased by 66%.

Consequently, the model requires more floating-point operations (FLOPs) and a longer inference time. This increase is expected, as ConvNeXt-Small is substantially larger and more computationally demanding than the ResNet34 backbone employed in RCFSNet, whereas the proposed scale-sensitive module contributes only a relatively small increase in model complexity. Despite the additional computational costs, RoadNeXt-CSS maintains an inference speed of 11.66 FPS on an NVIDIA A100 GPU with an input resolution of 1024 × 1024 pixels. Furthermore, the proposed model achieves superior road extraction performance, demonstrating a favourable trade-off between accuracy and computational efficiency.

5 FINDINGS AND DISCUSSION

The experimental results and paired *t*-test analysis collectively demonstrate that RoadNeXt-CSS achieves statistically significant and practically meaningful improvements over the baseline RCFSNet and several state-of-the-art road extraction models. On the Massachusetts Roads dataset, RCFSNet achieved an IoU of 0.6424, whereas the proposed RoadNeXt-CSS improved this value to 0.6779, representing an absolute gain of 3.55%. Similarly, on the DeepGlobe dataset, RoadNeXt-CSS attains an IoU of 0.6984 compared to 0.6606 for RCFSNet, corresponding to a 3.78% improvement. The paired *t*-test confirms that these improvements are statistically significant, indicating that the observed gains are not due to random variation across test samples.

Beyond that, it has been demonstrated that RoadNeXt-CSS has robust performance in various scenes, either urban or rural settings. Based on qualitative inspection, RoadNeXt-CSS produces more continuous road networks, recovers a larger number of narrow road segments and improves junction completeness compared with RCFSNet and other baseline methods. These characteristics are important for downstream applications that rely on accurate and complete road network representations, such as transportation planning, disaster response, navigation and urban monitoring. Therefore, the observed performance gains are not only statistically significant but also practically meaningful, as they translate into more reliable and operationally useful road extraction results in real-world scenarios.

Statistical analysis further clarifies the contribution of individual design choices. Replacing the ResNet34 encoder in RCFSNet with ConvNeXt-based encoders consistently results in significant performance gains on both datasets. This confirms that modern convolutional backbones provide stronger and more discriminative feature representations for thin and elongated road structures. However, comparisons between ConvNeXt-Tiny and ConvNeXt-Small do not consistently reach statistical significance, suggesting that encoder scaling alone yields diminishing returns once a sufficiently expressive backbone is employed.

Inverse-frequency-weighted (IFW) loss plays a dominant role in the Massachusetts Roads dataset, where the class imbalance between road and background pixels is severe. Models trained with IFW consistently outperform both the baseline and unweighted ConvNeXt variants, primarily through improved recall and reduced false negatives. In contrast, on the DeepGlobe dataset, the contribution of IFW becomes less pronounced once scale-sensitive decoding is applied, indicating a saturation effect when multiscale spatial information is already well captured.

Scale-sensitive decoding has a stronger impact on the DeepGlobe dataset, where road width, orientation and surrounding context vary substantially across regions. The paired *t*-test shows statistically significant improvements over RCFSNet when scale-sensitive feature weighting is introduced, confirming the importance of explicit multiscale modelling for geographically diverse imagery. While its effect on the Massachusetts Roads dataset is less consistent when used alone, scale-sensitive decoding does not degrade performance and provides complementary benefits when combined with IFW and a modern encoder.

Despite these improvements, several limitations remain. The most prominent shortcoming is the persistence of false positive predictions, particularly in visually complex regions where non-road surfaces share similar spectral or textural characteristics with roads, such as rooftops, parking areas and bare soil. This indicates that while RoadNeXt-CSS improves recall and overall IoU, further refinement is required to enhance precision and suppress spurious detections.

Additionally, the evaluation in this study relies on pixel-level metrics (IoU, F1 score, precision and recall), which are standard for the Massachusetts Roads and DeepGlobe benchmarks but do not directly capture road network connectivity. Topology-aware metrics, such as clDice or APLS, require either skeletonized or graph-structured

ground truth annotations that are not provided by these datasets, and their adoption would preclude direct comparison with existing methods evaluated under the same protocol.

Future work will focus on addressing these limitations through architectural, postprocessing and preprocessing strategies. More advanced encoder variants, such as ConvNeXtV2, will be explored to further strengthen feature discrimination and contextual understanding. Future evaluation will also incorporate connectivity-aware metrics such as clDice and APLS on datasets that provide the necessary graph-structured or skeletonized ground truth annotations to more comprehensively assess road network continuity and topological correctness.

In addition, lightweight postprocessing techniques, including connected component analysis and small-blob removal, will be investigated to reduce isolated false positives while preserving road continuity. Preprocessing methods, such as colour normalization and contrast enhancement, may also be incorporated to reduce appearance variability across scenes and improve robustness under diverse illumination and environmental conditions.

Overall, the results indicate that RoadNeXt-CSS represents a statistically validated and quantitatively superior improvement over RCFSNet and existing approaches, while also highlighting clear directions for future enhancement to improve reliability in real-world remote sensing applications.

5.1 Comparison with other deep learning-based methods

This subsection presents a performance comparison between the original RCFSNet, RoadNeXt-CSS and several established encoder-decoder-based deep learning models, including U-Net, LinkNet34, D-LinkNet50 and MECA-Net, on the Massachusetts Roads and DeepGlobe datasets. To ensure a fair evaluation, all the models assessed on the DeepGlobe dataset were retrained using the same dataset split, preprocessing pipeline, optimization settings, training schedule and evaluation protocol. Only the network architecture was changed between experiments. As shown in Table 8 for the Massachusetts Roads dataset and Table 9 for the DeepGlobe dataset, RoadNeXt-CSS consistently achieves the highest IoU and F1 score values, demonstrating a strong ability to detect road regions and outperforming all the competing methods in terms of IoU, recall and F1 score. Importantly, the performance improvement over RCFSNet as the most structurally comparable baseline is statistically validated, confirming that the proposed enhancements translate into measurable gains beyond architectural similarity. Nevertheless, the presence of false positives indicates that further refinement is required to improve precision, leading to higher IoU values.

To offer readers a clearer understanding of the proposed model and validate its effectiveness, we include representative output comparisons between our method and the baseline. We selected five representative output examples from the Massachusetts Roads dataset and five from the DeepGlobe dataset, where the input images are acquired from diverse landscapes, including rural, industrial, suburban and dense urban areas. For each example, we present the input image, the corresponding ground truth, the prediction images from U-Net, D-LinkNet50, the baseline (RCFSNet) and the proposed model (RoadNeXt-CSS). In all the outputs, green denotes true positives, red indicates false negatives and blue represents false positives with respect to the ground truth. A brief comparative analysis is provided in the paragraph following it. The prediction comparison with the Massachusetts Roads and DeepGlobe datasets is shown in Figure 7 and Figure 8, respectively.

Table 8. *Performance comparison of RoadNeXt-CSS and other deep learning-based methods on Massachusetts Roads dataset.*

Methods	IoU	Precision	Recall	F1 score
U-Net (Ronneberger et al., 2015)	0.6610	**0.8797**	0.7259	0.7937
LinkNet34 (Chaurasia & Culurciello, 2017)	0.6406	0.8637	0.7127	0.7780
D-LinkNet50 (L. Zhou et al., 2018)	0.6459	0.8741	0.7119	0.7823
MACU-Net (R. Li et al., 2022)	0.6457	0.8661	0.7165	0.7824
MECA-Net (Jie et al., 2022)	0.6582	0.8063	0.7819	0.7939
RCFSNet(Yang et al., 2023)	0.6424	0.8714	0.7095	0.7788
RoadNeXt-CSS	**0.6779**	0.8105	**0.8047**	**0.8061**

Table 9. *Performance comparison of RoadNeXt-CSS and other deep learning-based methods on DeepGlobe dataset.*

Methods	IoU	Precision	Recall	F1 score
U-Net (Ronneberger et al., 2015)	0.5980	0.8104	0.7072	0.7325
LinkNet34 (Chaurasia & Culurciello, 2017)	0.6350	**0.8131**	0.7428	0.7635
D-LinkNet50 (L. Zhou et al., 2018)	0.6569	0.7749	0.8176	0.7836
MACU-Net (R. Li et al., 2022)	0.6251	0.7915	0.7515	0.7574
RCFSNet(Yang et al., 2023)	0.6606	0.7856	0.8097	0.7860
RoadNeXt-CSS	**0.6984**	0.7871	**0.8629**	**0.8146**

On Massachusetts Roads (Figure 7), U-Net demonstrates slightly stronger performance than D-LinkNet50 and RCFSNet. It more consistently preserves long road segments with fewer interruptions, particularly in open and low-density regions, resulting in reduced false negatives along primary roads. In contrast, D-LinkNet50 exhibits increased fragmentation and occasional false positives near complex boundaries, while RCFSNet shows missed detections on elongated or weakly textured roads, indicating reduced robustness when the visual complexity is lower, but road continuity is dominant.

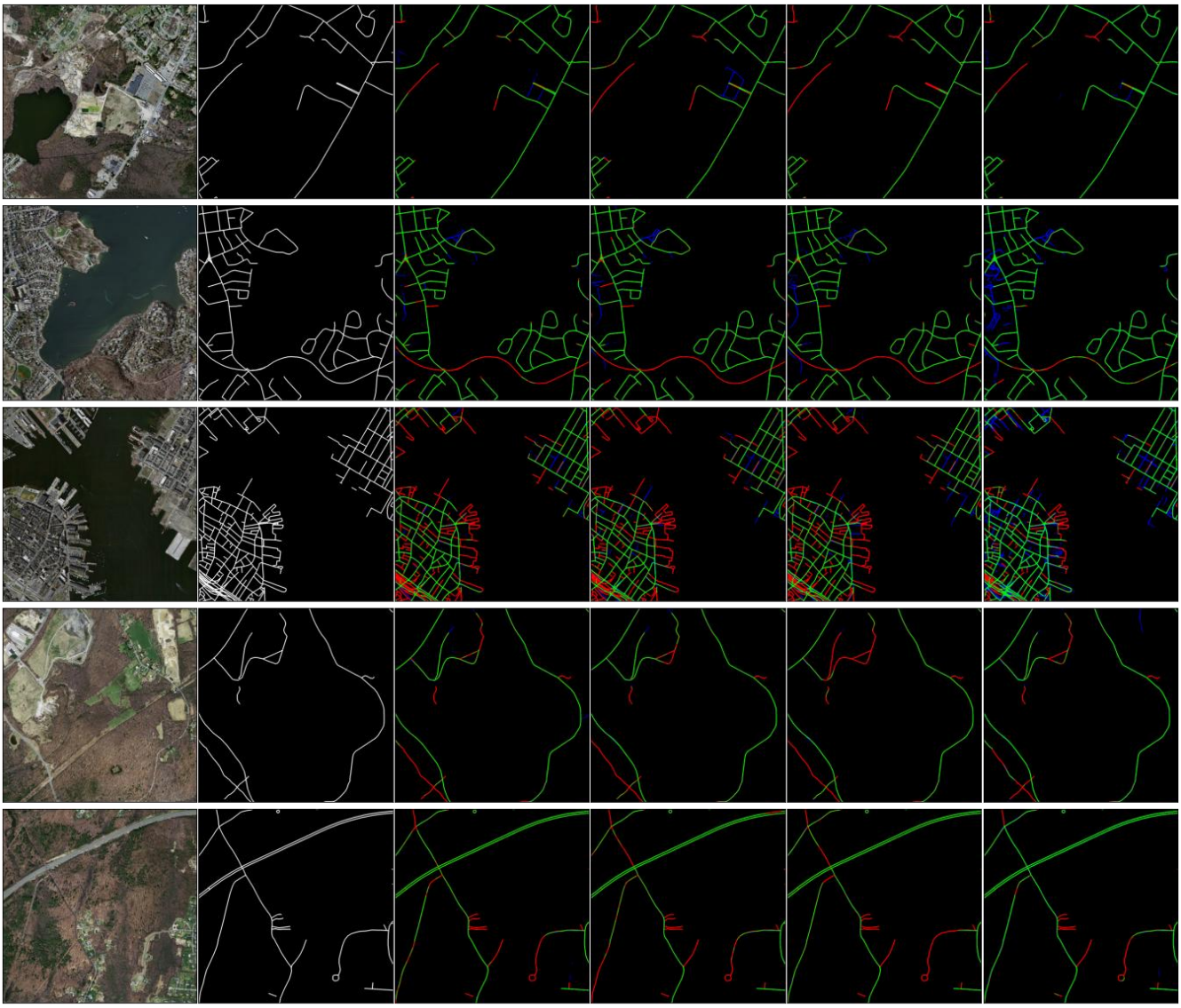


Figure 7. *Example of road extraction result from remote sensing images in Massachusetts Road dataset. From the top to bottom rows, the raw input images are: (1) 23428540.tif, (2) 25079170.tif, (3) 23729020.tif, (4) 18478900.tif and (5) 10378780.tif.*

Despite these variations among baseline methods, all three struggle to accurately capture thin or branching roads, especially in suburban neighbourhoods and curved coastal layouts. False negatives remain prevalent at junctions and on narrow residential streets, whereas false positives occasionally appear in visually similar non-road regions, such as shorelines or cleared land. These observations suggest that architectural depth or dilated convolution alone is insufficient to address scale variation and class imbalance inherent in road extraction tasks.

RoadNeXt-CSS consistently delivers complete and structurally faithful predictions across all the examples. Its robust feature extractor and imbalance-aware learning substantially reduce false negatives, even in scenes dominated by background pixels. Although, on the other hand, false positives increase slightly. Moreover, the scale-sensitive decoder enables effective modelling of roads with varying widths, allowing the proposed method to recover narrow residential roads and subtle connectors that are missed by other approaches. The dominance of true positives and improved network connectivity visually confirms the superior generalization capability of RoadNeXt-CSS on the Massachusetts Roads dataset, complementing the quantitative performance gains.

On DeepGlobe (Figure 8), U-Net shows the most pronounced limitations, with predictions that are highly fragmented and dominated by false negatives, particularly for narrow roads and complex intersections. Many secondary and residential roads are either partially detected or completely missed, reflecting insufficient feature discrimination between road and background regions. D-LinkNet50 improves upon U-Net by capturing longer road segments and reducing fragmentation; however, it still exhibits notable false negatives at intersections and along thin or curved roads, indicating limited robustness to scale variation and class imbalance.

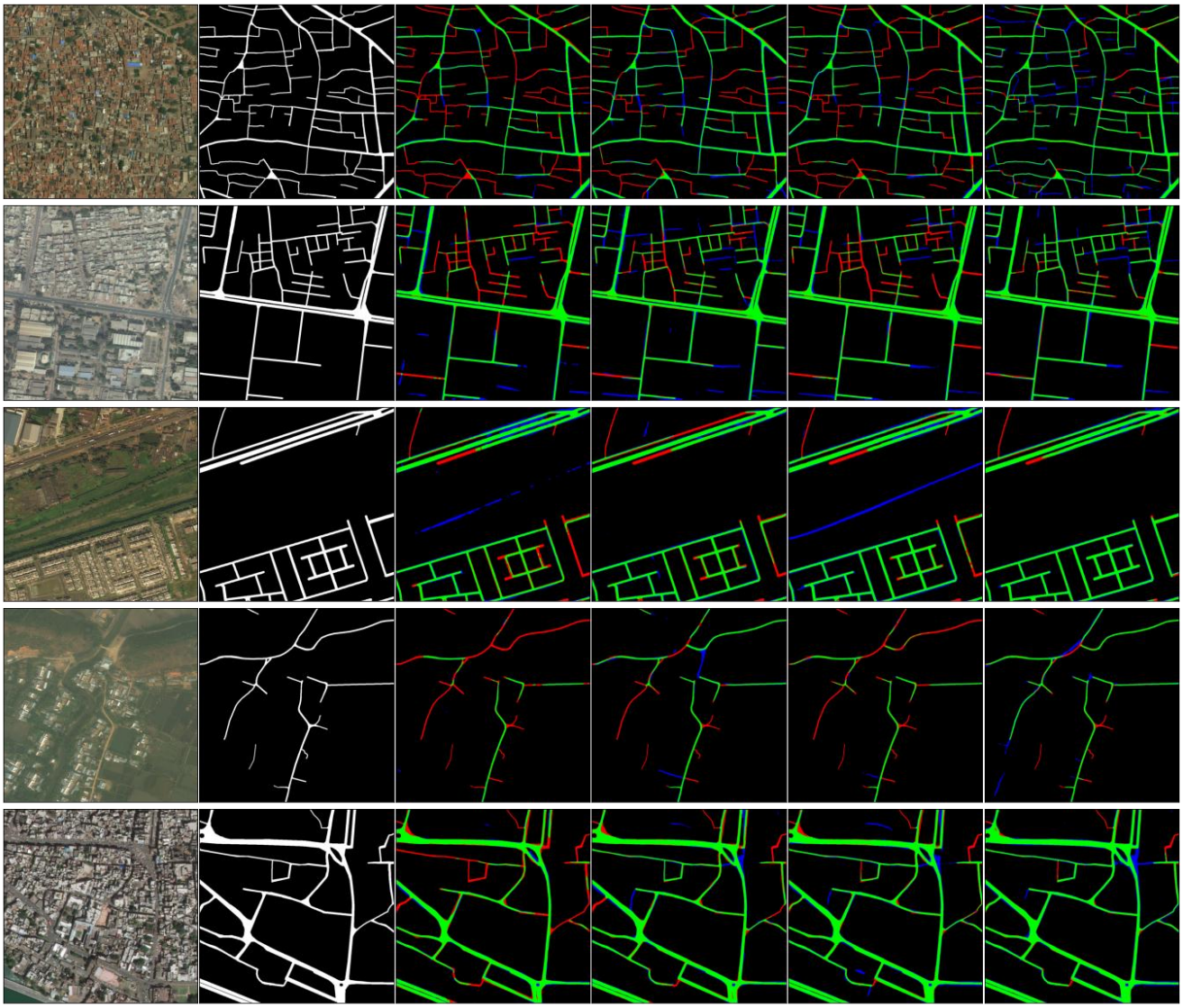


Figure 8. *Example of road extraction result from remote sensing images in DeepGlobe dataset. From the top to bottom rows, the raw input images are: (1) 22415.jpg, (2) 442978.jpg, (3) 604916.jpg, (4) 956106.jpg and (5) 779760.jpg.*

RCFSNet, as the most structurally comparable baseline, also enhances continuity by better modelling long-range dependencies and dominant road structures. Major roads and primary connections are generally well recovered, but the model continues to underperform in fine-grained details. This is visible in the persistence of false negatives on narrow streets and minor branches, as well as occasional false positives in ambiguous background areas, suggesting sensitivity to scale and residual imbalance between road and non-road pixels. The most visible false positive of RCFSNet is displayed in the third image, where it mispredicted the waterway as a road, showing that its feature extraction capabilities are not very strong.

In contrast, RoadNeXt-CSS consistently produces complete and topologically coherent road networks. Its robust feature extractor, coupled with imbalance-aware learning, suppresses false negatives by strengthening the representation of minority road pixels. In addition, the scale-sensitive decoder enables adaptive handling of roads with varying widths, allowing the model to capture narrow residential streets and subtle connectors that other methods miss. The dominance of true positives and improved junction completeness across all scenes visually confirms the superiority of RoadNeXt-CSS, aligning with the quantitative gains reported in the experimental results.

6 CONCLUSION

This study presented RoadNeXt-CSS, a tool for road extraction from remote sensing imagery built on the RCFSNet encoder-decoder structure. Three key contributions were introduced: a ConvNeXt-based encoder for improved feature discrimination between roads and surrounding areas, a scale-sensitive (ScaleSen) decoder refinement module to emphasize important spatial features while reducing background noise and an inverse frequency weighting (IFW) training strategy to mitigate class imbalance. Notably, the IFW **pos_weight** parameter is derived from the pixelwise class frequency distribution of the training set, meaning that its optimal value is inherently dataset-dependent. While this allows the loss calibration to adapt to the specific imbalance characteristics of a given dataset, practitioners applying RoadNeXt-CSS to new datasets should recalibrate this parameter accordingly to ensure effective training behaviour.

The experimental results directly address the four research questions posed in this study: RQ1 is affirmed by consistent IoU gains from the ConvNeXt encoder over the ResNet34 baseline; RQ2 is supported by improved recall attributable to the IFW loss; RQ3 is confirmed by the contribution of the ScaleSen decoder to multi-scale road detection, particularly when combined with IFW; and RQ4 is validated by the paired *t*-test results demonstrating that the three components together provide statistically significant and complementary improvements across both benchmark datasets.

The experimental results demonstrate that the proposed model outperforms the baseline and several other encoder-decoder architectures in terms of intersection over union (IoU), recall and F1 scores, yielding more accurate road area detections. It achieves an IoU of 0.6779 on Massachusetts Roads and 0.6984 on DeepGlobe, as well as F1 scores of 0.8061 on Massachusetts Roads and 0.8146 on DeepGlobe. These findings underscore the effectiveness of our integrated approach in addressing common challenges in road extraction tasks. The paired *t*-test analysis confirms that the three improvements proposed in RoadNeXt-CSS are statistically significant. The ConvNeXt-based encoder provides consistent gains across datasets, whereas the other two (weighted loss function and scale-sensitive decoder) act as complements to further enhance the road segmentation performance.

Qualitative inspection suggests that the improvements provided by RoadNeXt-CSS are particularly beneficial in visually challenging regions, including thin road segments, partially occluded roads and areas with substantial variation in road width and surrounding context. The IFW loss improves the sensitivity to minority road pixels, whereas the scale-sensitive decoder enhances the integration of multiscale features. These results highlight the fact that road extraction success is built on the combination of feature representation, optimization behaviour and multiscale information integration. Since these three areas function as orthogonal improvements, our study advocates a comprehensive design approach that targets each level simultaneously. This indicates that future models should prioritize integrated frameworks that make use of these distinct benefits rather than relying on isolated modifications to architecture. Nevertheless, some errors persist in regions containing elongated non-road structures with road-like appearances, such as drainage channels, artificial boundaries and visually plausible road continuations. The model may also overestimate road continuity beneath dense vegetation, indicating that future work should further improve discrimination between roads and visually similar linear objects.

In future experiments, we will address the high false positive issue with techniques such as a better encoder and the postprocessing module to remove small blobs. Preprocessing methods, such as colour enhancement, could also be employed. Future work will also evaluate RoadNeXt-CSS on additional road extraction datasets and investigate cross-dataset generalization to assess robustness in respect of different geographical regions, imaging conditions and annotation standards. Nevertheless, the more accurate road extraction achieved by RoadNeXt-CSS enables better-informed urban planning decisions, strengthens disaster management and risk mitigation strategies, supports more reliable transportation analysis and improves the overall effectiveness of geospatial intelligence systems.

ADDITIONAL INFORMATION AND DECLARATIONS

Acknowledgments: We would like to thank the Center for Computation, the National Agency for Research and Innovation (BRIN) for providing the computational resources needed to carry out this work, and Tanzi Mubaroq Santoso (BRIN) for drawing the figures of the FSFF, MSCE, and RoadNeXt-CSS architecture. We also thank all researchers in the Computer Vision and Image Processing Research Group, Research Center for Artificial Intelligence and Cybersecurity, the National Research and Innovation Agency, Republic of Indonesia (BRIN), Bandung, and South Tangerang, Indonesia for their support and cooperation in this research.

Conflict of Interests: The authors declare no conflict of interest.

Author Contributions: R.A.P.: Conceptualization, Methodology, Software, Formal analysis, Investigation, Resources, Data curation, Writing – Original draft preparation. A.M.A.: Conceptualization, Validation, Supervision, Writing – Reviewing and Editing, Project administration. R.R.: Methodology, Writing – Reviewing and Editing.

Statement on the Use of Artificial Intelligence Tools: During the preparation of this manuscript, the authors used ChatGPT and Claude free version to assist with (1) refining the language and phrasing of sentences drafted by the authors, and (2) checking the internal consistency and validity of the authors' own analysis. The tool's output was used only as an aid for review; all analytical decisions, interpretations, and conclusions remain those of the authors, who reviewed, verified, and take full responsibility for all content in the manuscript.

Data Availability: The data supporting the findings are openly available in the GitHub repository https://github.com/WaOnEmperoR/roadnext_css.

REFERENCES

Ahmed, Md. I., Foysal, Md., Chaity, M. D., & Hossain, A. B. M. A. (2024). DeepRoadNet: A deep residual based segmentation network for road map detection from remote aerial image. *IET Image Processing*, 18(1), 265–279. https://doi.org/10.1049/ipr2.12948

Akhtarmanesh, A., Abbasi-Moghadam, D., Sharifi, A., Yadkouri, M. H., Tariq, A., & Lu, L. (2024). Road Extraction from Satellite Images Using Attention-Assisted UNet. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17, 1126–1136. https://doi.org/10.1109/JSTARS.2023.3336924

Anil, P., & Natarajan, S. (2010). A novel approach using active contour model for semi-automatic road extraction from high resolution satellite imagery. In *2010 Second International Conference on Machine Learning and Computing*, (pp. 263–266). IEEE. https://doi.org/10.1109/ICMLC.2010.36

Chaurasia, A., & Culurciello, E. (2017). LinkNet: Exploiting encoder representations for efficient semantic segmentation. In *2017 IEEE Visual Communications and Image Processing (VCIP)*, (pp. 1–4). IEEE. https://doi.org/10.1109/VCIP.2017.8305148

Dai, L., Zhang, G., & Zhang, R. (2023). RADANet: Road Augmented Deformable Attention Network for Road Extraction From Complex High-Resolution Remote-Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1–13. https://doi.org/10.1109/TGRS.2023.3237561

Du, Y., Sheng, Q., Zhang, W., Zhu, C., Li, J., & Wang, B. (2023). From local context-aware to non-local: A road extraction network via guidance of multi-spectral image. *ISPRS Journal of Photogrammetry and Remote Sensing*, 203, 230–245. https://doi.org/https://doi.org/10.1016/j.isprsjprs.2023.07.026

Huan, H., Sheng, Y., Zhang, Y., & Liu, Y. (2022). Strip Attention Networks for Road Extraction. *Remote Sensing*, 14(18). https://doi.org/10.3390/rs14184516

Jie, Y., He, H., Xing, K., Yue, A., Tan, W., Yue, C., Jiang, C., & Chen, X. (2022). MECA-Net: A MultiScale Feature Encoding and Long-Range Context-Aware Network for Road Extraction from Remote Sensing Images. *Remote Sensing*, 14, 5342. https://doi.org/10.3390/rs14215342

Li, R., Duan, C., Zheng, S., Zhang, C., & Atkinson, P. M. (2022). MACU-Net for Semantic Segmentation of Fine-Resolution Remotely Sensed Images. *IEEE Geoscience and Remote Sensing Letters*, 19, 1–5. https://doi.org/10.1109/LGRS.2021.3052886

Li, Y., Zhang, R., & Wu, Y. (2017). Road network extraction in high-resolution SAR images based CNN features. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, (pp. 1664–1667). IEEE. https://doi.org/10.1109/IGARSS.2017.8127293

Liu, R., Wu, J., Lu, W., Miao, Q., Zhang, H., Liu, X., Lu, Z., & Li, L. (2024). A Review of Deep Learning-Based Methods for Road Extraction from High-Resolution Remote Sensing Images. *Remote Sensing*, 16(12), 2056. https://doi.org/10.3390/rs16122056

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A ConvNet for the 2020s. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 11966–11976). IEEE. https://doi.org/10.1109/CVPR52688.2022.01167

Mei, J., Li, R.-J., Gao, W., & Cheng, M.-M. (2021). CoANet: Connectivity Attention Network for Road Extraction From Satellite Imagery. *IEEE Transactions on Image Processing*, 30, 8540–8552. https://doi.org/10.1109/TIP.2021.3117076

Mo, S., Shi, Y., Yuan, Q., & Li, M. (2024). A Survey of Deep Learning Road Extraction Algorithms Using High-Resolution Remote Sensing Images. *Sensors*, 24(5), 1708. https://doi.org/10.3390/s24051708

Perdana, R. A., Arimurthy, A. M., & Risnandar. (2024). Remote Sensing Scene Classification using ConvNeXt-Tiny Model with Attention Mechanism and Label Smoothing. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, 8(3), 389–400. https://doi.org/10.29207/resti.v8i3.5731

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (pp. 234–241). Springer. https://doi.org/10.1007/978-3-319-24574-4_28

Simler, C. (2011). An improved road and building detector on VHR images. In 2011 IEEE International Geoscience and Remote Sensing Symposium, (pp. 507–510). IEEE. https://doi.org/10.1109/IGARSS.2011.6049176

Soni, Y., Meena, U., Mishra, V. K., & Soni, P. K. (2025). AM-UNet: Road Network Extraction from high-resolution Aerial Imagery Using Attention-Based Convolutional Neural Network. *Journal of the Indian Society of Remote Sensing*, 53(1), 135–147. https://doi.org/10.1007/s12524-024-01974-3

Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., & Jorge Cardoso, M. (2017). Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, (pp. 240–248). Springer. https://doi.org/10.1007/978-3-319-67558-9_28

Sugino, T., Kawase, T., Onogi, S., Kin, T., Saito, N., & Nakajima, Y. (2021). Loss Weightings for Improving Imbalanced Brain Structure Segmentation Using Fully Convolutional Networks. *Healthcare*, 9(8), 938. https://doi.org/10.3390/healthcare9080938

Tan, X., Xiao, Z., Wan, Q., & Shao, W. (2021). Scale Sensitive Neural Network for Road Segmentation in High-Resolution Remote Sensing Images. *IEEE Geoscience and Remote Sensing Letters*, 18(3), 533–537. https://doi.org/10.1109/LGRS.2020.2976551

Wang, H., Bai, L., Ji, B., Shen, J., Jin, Y., Ye, F., & Zhu, C. (2023). Gated Fusion and Dual Attention Network for Road Extraction from Satellite Imagery. In *2023 8th International Conference on Image, Vision and Computing (ICIVC),* (pp. 495–500). IEEE. https://doi.org/10.1109/ICIVC58118.2023.10270019

Wang, Q., Bai, H., He, C., & Cheng, J. (2022). Fe-LinkNet: Enhanced D-LinkNet with Attention and Dense Connection for Road Extraction in High-Resolution Remote Sensing Images. In *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, (pp. 3043–3046). IEEE. https://doi.org/10.1109/IGARSS46834.2022.9883026

Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., & Hu, Q. (2020). ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* (pp. 11531–11539). IEEE. https://doi.org/10.1109/CVPR42600.2020.01155

Wang, W., Yang, N., Zhang, Y., Wang, F., Cao, T., & Eklund, P. (2016). A Review of Road Extraction from Remote Sensing Images. *Journal of Traffic and Transportation Engineering*, 3(3), 271–282. https://doi.org/10.1016/j.jtte.2016.05.005

Wang, X., Hu, Z., Shi, S., Hou, M., Xu, L., & Zhang, X. (2023). A deep learning method for optimizing semantic segmentation accuracy of remote sensing images based on improved UNet. *Scientific Reports*, 13(1), 7600. https://doi.org/10.1038/s41598-023-34379-2

Wang, X., Qin, C., Bai, M., Ma, Q., & Li, G. (2025). CAFormer: A connectivity-aware vision transformer for road extraction from remote sensing images. *The Visual Computer*, 41(10), 7965–7981. https://doi.org/10.1007/s00371-025-03849-1

Wei, D., Li, P., Xie, H., & Xu, Y. (2023). DRCNet: Road Extraction From Remote Sensing Images Using DenseNet With Recurrent Criss-Cross Attention and Convolutional Block Attention Module. *IEEE Access*, 11, 126879–126891. https://doi.org/10.1109/ACCESS.2023.3332121

Wei, Z., & Zhang, Z. (2023). Remote Sensing Image Road Extraction Network Based on MSPFE-Net. *Electronics*, 12(7), 1713. https://doi.org/10.3390/electronics12071713

Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). CBAM: Convolutional Block Attention Module. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Computer Vision – ECCV 2018* (pp. 3–19). Springer. https://doi.org/10.1007/978-3-030-01234-2_1

Yang, Z., Zhou, D., Yang, Y., Zhang, J., & Chen, Z. (2023). Road Extraction From Satellite Imagery by Road Context and Full-Stage Feature. *IEEE Geoscience and Remote Sensing Letters*, 20, 1–5. https://doi.org/10.1109/LGRS.2022.3228967

Zhang, C., Murai, S., & Baltsavias, E. P. (1999). Road network detection by mathematical morphology. In *ISPRS Workshop" 3D Geospatial Data Production: Meeting Application Requirements*, (pp. 1–18). ISPRS.

Zhang, P., Niu, L., Cai, M., Chen, H., & Sun, X. (2024). AAUConvNeXt: Enhancing Crop Lodging Segmentation with Optimized Deep Learning Architectures. *Plant Phenomics*, 6, 0182. https://doi.org/10.34133/plantphenomics.0182

Zhang, Y., Zhang, L., Niu, P., & Xu, W. (2023). AGF-Net: Adaptive Road Extraction Network with Global Feature Fusion. In *2023 5th International Conference on Data-Driven Optimization of Complex Systems*, (pp. 1–7). IEEE. https://doi.org/10.1109/DOCS60977.2023.10294802

Zhang, Z., Sun, X., & Liu, Y. (2022). GMR-Net: Road-Extraction Network Based on Fusion of Local and Global Information. *Remote Sensing*, 14(21), 5476. https://doi.org/10.3390/rs14215476

Zhou, J., Bischof, W. F., & Caelli, T. (2006). Road tracking in aerial images based on human–computer interaction and Bayesian filtering. *ISPRS Journal of Photogrammetry and Remote Sensing*, 61(2), 108–124. https://doi.org/https://doi.org/10.1016/j.isprsjprs.2006.09.002

Zhou, L., Zhang, C., & Wu, M. (2018). D-LinkNet: LinkNet with Pretrained Encoder and Dilated Convolution for High Resolution Satellite Imagery Road Extraction. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),* (pp. 182–186). IEEE. https://doi.org/10.1109/CVPRW.2018.00034

Acta Informatica Pragensia is published by the Prague University of Economics and Business, Czech Republic | eISSN: 1805-4951