frontiers | Frontiers in Remote Sensing

TYPE Original Research
PUBLISHED 26 September 2025
DOI 10.3389/frsen.2025.1668978

OPEN ACCESS

EDITED BY
Rui Li,
University of Warwick, United Kingdom

REVIEWED BY
Miguel Angel Manso Callejo,
Polytechnic University of Madrid, Spain
Liang Huang,
Kunming University of Science and Technology, China

*CORRESPONDENCE
Qi Feng,
fengqi@ppsuc.edu.cn

RECEIVED 18 July 2025
ACCEPTED 15 August 2025
PUBLISHED 26 September 2025

CITATION
Cui S, Feng Q, Ji L, Liu X and Guo B (2025) HPLNet: a hierarchical perception lightweight network for road extraction.
Front. Remote Sens. 6:1668978.
doi: 10.3389/frsen.2025.1668978

HPLNet: a hierarchical perception lightweight network for road extraction

Shilin Cui[1], Qi Feng[1]*, Luyan Ji[2,3,4], Xiaowen Liu[1] and Baofeng Guo[1]

[1]Department of Information and Cyber Security, People's Public Security University of China, Beijing, China, [2]Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China, [3]School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing, China, [4]Key Laboratory of Target Cognition and Application Technology, Chinese Academy of Sciences, Beijing, China

With the progression of remote sensing technologies, extracting road networks from satellite imagery has emerged as a pivotal research domain in both Geographic Information Systems and Intelligent Transportation Systems. Recognizing the difficulty in balancing lightweight network design with extraction accuracy, the challenge of synergistically preserving global road connectivity and local details, and the hardship in effectively integrating low-level features with high-level representations to achieve full coupling between road details and semantic understanding in road extraction from remote sensing images, this study introduces a Hierarchical Perception Lightweight Network for road extraction (HPLNet). This innovative network integrates shallow perception part and deep perception part, aiming to optimize the trade-off between inference efficiency and extraction accuracy. In shallow perception, directional stripe convolutions capture road details, while deep perception integrates a spatial-channel semantic awareness network to bridge local and global information, boosting road semantic feature extraction. Moreover, to extend the model's reception at both pixel and semantic levels, each network component strategically introduces parameter-free channel shift operations. HPLNet attains state-of-the-art efficiency in balancing parameter footprint and inference latency: its parameter count is merely 22% of that of U-Net, while its inference speed is 18% faster than FCN. Concurrently, it delivers competitive segmentation metrics on the Massachusetts dataset, achieving an IoU of 64.32% and an F1 score of 79.96%. Experimental results demonstrate that the proposed network achieves superior performance in both segmentation accuracy and model complexity, thereby offering an efficient solution for real-time deployment on edge devices.

KEYWORDS

road extraction, satellite imagery, hierarchical perception, lightweight network, channel shift

1 Introduction

Extracting roads from high-resolution remote sensing images is of significant practical value across various domains, including map delineation, urban planning Qian et al. (2021); Liu and Wang (2011); Qi et al. (2020), traffic monitoring Cruz et al. (2022); Shao et al. (2023); Seid et al. (2020), environmental monitoring Xu et al. (2018); Wan et al. (2019); Dong (2012), and disaster response Wu et al. (2018); Huang et al. (2021); Wang et al. (2015).

In resource-constrained environments, lightweight network models enable more efficient operations by reducing computational resource consumption, thus enhancing the efficiency and real-time performance of road extraction. This holds critical significance for rapid road information acquisition in practical applications, particularly for enabling timely and accurate road network extraction under scenarios with limited computational capabilities Zhou et al. (2021).

In the early stages of road extraction research in remote sensing imagery, the research methodology focuses on morphological based methods and machine learning methods to extract roads Liu et al. (2015); Wang et al. (2016); Lian et al. (2020a); Yuan et al. (2021), these methods are limited to single-channel grayscale images, on the one hand, limited by the fixed arithmetic scale, it is difficult to cope with the real-time processing demand in complex scenes; on the other hand, in the process of multi-scale road feature resolution, due to the lack of hierarchical computation optimization mechanism, resulting in exponential degradation of the efficiency of feature extraction, and at the same time the intervention of manual annotation further aggravates the risk of error accumulation. Deep learning has led to the development of semantic segmentation neural networks using deep convolution, which perform well across various image tasks. These methods typically use an encoder-decoder structure Badrinarayanan et al. (2017). The encoder extracts image features layer by layer, and the decoder integrates these features for accurate pixel classification. Convolutional neural networks (CNNs) like those in Ronneberger et al. (2015a); Badrinarayanan et al. (2017); Chaurasia and Culurciello (2017); Chen et al. (2018), are effective for image semantic segmentation. U-Net Ronneberger et al. (2015a) combines a contracting and an expansive path to extract context and location information, restoring image details through upsampling. However, CNN-based methods suffer from inefficient, affected by shadows, and sensitive to occlusions, leading to poor continuity and extensibility in road extraction. To enhance the quality of road network topology, methods based on graph structures are receiving increasing attention. Graph Convolutional Networks (GCNs), initially applied to knowledge graphs, have recently been employed for feature extraction from natural images. Cui et al. (2021) proposed a novel road extraction method combining superpixel segmentation and GCN, which retains more spatial detail information and effectively improves the completeness of the extracted roads. However, graph-based segmentation methods are computationally complex and parameterized, and their effectiveness is often affected by the quality of the initial road segmentation results. In recent years, the remarkable achievements of the Transformer architecture in the realm of natural language processing have spurred its adoption in computer vision tasks, including image classification and segmentation. The Transformer model utilizes the self-attention mechanism and positional coding, which is well adapted to the connectivity and continuity characteristics of road networks, thus achieving excellent performance in tasks such as image segmentation and feature extraction. RoadViT integrates MobileViT for encoding advanced contextual information, employs a pyramid decoder for merging features across multiple scales, and enhances the quality of remote sensing images through data augmentation techniques. However, these methods currently face the following challenges.

1. In resource-constrained scenarios for road extraction from remote sensing images, the design of a lightweight network is essential. However, the challenge remains in maintaining extraction accuracy under strict lightweight constraints, presenting a critical trade-off between model efficiency and performance in practical remote sensing applications.
2. The long and narrow coverage of roads spans large areas, making them susceptible to influences from shadows and tree occlusions. Maintaining the lightweight design and accuracy of road extraction networks, the challenge remains in how to preserve both the global connectivity of roads and their excellent local details during extraction.
3. Decomposing road fine-grained details and high-level semantic understanding, and fully coupling these two remain a challenge. The key lies in balancing the extraction of pixel-level details and contextual semantic information, which demands an effective mechanism to integrate low-level features with high-level representations in remote sensing road extraction.

To address the challenges, a hierarchical perception lightweight network for road extraction (HPLNet) is proposed here. As illustrated in Figure 1a, we have developed a lightweight network design that efficiently optimizes the parameter count, achieves competitive FLOPs, and preserves robust road extraction accuracy. As illustrated in Figure 1c, we adopt a hierarchical perception strategy, employing a convolution-based feature extraction strategy to extract Fea_s when pixel features are abundant, and a lightweight attention mechanism-based feature extraction strategy Fea_D when semantic features are prominent. As illustrated in Figure 1b, to address the challenges of road extraction accuracy, global connectivity, and local detail preservation, we utilize the layer-wise extracted features Fea_s and Fea_D to guide local detail and global connectivity, respectively. This approach enables excellent extraction performance in terms of accuracy, topological continuity, and fine-grained detail.

The contributions of this paper are outlined as follows.

- We propose a lightweight road network that captures raw pixel details through shallow perception and obtains semantic information through deep perception, thereby achieving a balance between lightweight network design and feature extraction accuracy.
- We introduce a lightweight attention mechanism to collect long-distance road information and combine it with striped convolution to restore local road details, cleverly solving the difficult problem of balancing global connectivity and local details.
- We propose a parameter-free channel shift operation that achieves sufficient feature extraction at both deep and shallow perceptions, thereby achieving deep information coupling.

2 Related work

2.1 Road segmentation networks

Traditional road extraction methods are categorized into semi-automatic and fully automatic types Lian et al. (2020b). Niu (2006)

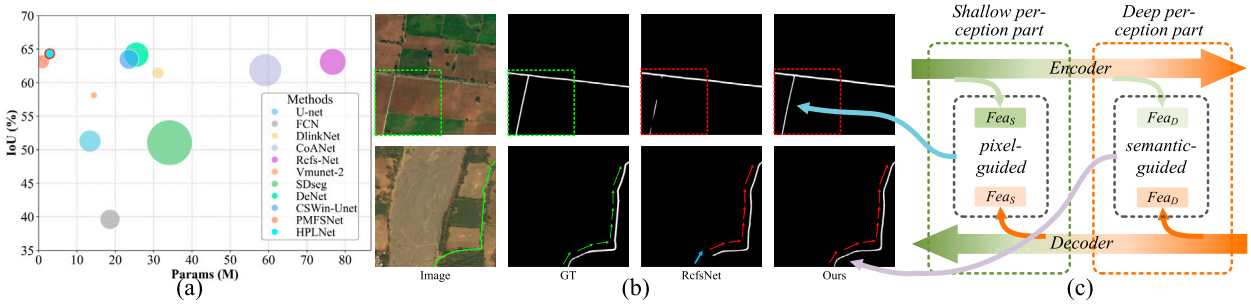

FIGURE 1
The Motivation of Our Network Design. **(a)** Comparison of the segmentation results of our method with other models on the Massachusetts dataset: the test image size is 1024 × 1024 pixels and the area of each bubble is proportional to the FLOPs(floating-point operations) of the corresponding model. **(b)** From left to right: satellite image, ground truth, prediction of RcfsNet and prediction of our method. **(c)** Network design. Blue arrow: Shallow features serve as pixel guidelines to guide the capture of road connectivity features through the shallow perception part. Purple arrow: Deep features serve as semantic guidelines to guide the capture of road extensibility features via the deep perception part.

presented a method integrating boundary gradients. It uses a geometric deformable model to minimize an objective function related to optimization problems based on road contours, linking the optimization problem to the propagation process of regular curves to address topological changes during curve deformation. However, this approach requires manual intervention, reducing work efficiency. Later, Yager and Sowmya (2003) employed SVM methods to extract road features from remote sensing images, but with relatively low accuracy. With the advancement of deep learning technology, a plethora of research efforts have developed segmentation networks using deep learning techniques, aiming to identify and extract roads from satellite imagery. Mnih and Hinton (2010) pioneered the integration of CNN models for road extraction in remote sensing images. Saito et al. (2016) employed a CNN in conjunction with channel-wise suppressed softmax to effectively train the network for feature extraction and prediction, yielding information about buildings and roads. In 2018, Zhou et al. (2018) introduced DlinkNet, which augmented the receptive field while maintaining high-resolution feature maps by incorporating cascaded dilated convolutions into the LinkNet architecture. The follow-up DlinkNetPlus Li et al. (2019) further optimized the model by reducing parameters and enhancing extraction precision. Mei et al. (2021) proposed CoANet, which addressed road connectivity and adaptability to road shapes by integrating a connection attention module with a strip convolution module, thereby mitigating the effects of occlusions. However, these networks often fall short in fully accounting for both global and local information during image processing, leading to limitations in segmentation accuracy and the retention of fine details. Li et al. (2021) presents DC-Net, a method fusing hollow convolution and ASPP. Its encoder-decoder structure, combined with multi-scale feature extraction, enables superior accuracy, enhanced connectivity, and improved occlusion resistance. Cheng et al. (2025) proposed a cascaded efficient road extraction network (CE-RoadNet) that combines smooth null convolutional residual blocks and introduces an attention-guided feature fusion (AGFF) sub-module for dynamically fusing features at different levels. Zhu et al. (2024) based on Swin Transformer, Wang designed Spatial Self-Attention (SSA) and Spatial MLP (SMLP) modules to extract road feature information more effectively. Zhou et al. (2021) proposed separable GCNs. two GCNs are used to capture global contextual information in space and channel to enhance the representation of road features. The road networks within remote sensing imagery are expansive yet intricate, a challenge that traditional methods struggle to address in terms of capturing global context and precise localization. The Transformer architecture, with its innovative self-attention mechanism, has proven effective in harnessing comprehensive information and contextual cues, facilitating parallel processing. Yang and colleagues Yang et al. (2022) leveraged the Swin Transformer for road extraction in remote sensing images, refining context acquisition and devising a supplementary module for foreground context information to bolster the interpretation of indistinct road segments. Wu et al. (2022) propose an improved semantic segmentation method based on ResNet. By introducing coordinate convolutions before and after encoding to enhance spatial edge information and a global information enhancement module to improve context perception, we address the problems of spatial feature loss, detail loss, and mis-extraction caused by convolutional pooling operations in deep learning road extraction. Yang and associates Yang Z. et al. (2023) introduced the RcfsNet, which capitalizes on multi-scale context extraction and a full-stage feature fusion module to elevate the quality of road segmentation in satellite images.

Deep learning-based road extraction methods have made significant progress in improving extraction accuracy, yet they still suffer from limitations in comprehensively integrating global contextual information and local fine-grained details. This fundamental gap leads to compromised segmentation precision and insufficient detail preservation, particularly in complex scenes with occlusions or topological discontinuities. Furthermore, approaches like DLinkNet and CoANet often achieve performance gains by increasing network depth and architectural complexity, which inevitably results in a steep rise in parameter counts and computational overhead. Similarly, Transformer-based models, despite their strong global modeling capabilities, impose substantial computational burdens due to their quadratic complexity in sequence length, making them infeasible for real-time applications in resource-constrained environments.

2.2 Coupling local and global information

In the realm of remote sensing imagery for road extraction, extensive research has explored the fusion of CNNs and Transformers to derive more comprehensive features. This hybrid approach leverages the spatial information capturing prowess of CNNs and the sequential data processing and long-range dependency handling capabilities of Transformers. Lu et al. (2020) designed a globally aware deep network (GAN) for road detection that includes a spatial awareness module (SAM) and a channel awareness module (CAM). Jamali et al. (2024) combined residual learning with UNet and ViT in ResUNetFormer while using Neighbourhood Attention Transformer for local feature enhancement. Wang and team Wang R. et al. (2023) fused CNN and Transformer architectures within the UNet framework, incorporating a dual upsampling module to enhance feature extraction and overall performance. Wang C. et al. (2023) advanced road extraction capabilities in remote sensing imagery by integrating Transformer-based ESTM and GDEM for context modeling, along with introducing the REF loss function in conjunction with a hybrid self-attention mechanism. Zhang et al. (2023b) introduced a Transformer-centric technique incorporating modules for intricate road feature extraction and the integration of global and local contexts. Yang Z.-X. et al. (2023) developed SSEANet, a novel framework that simultaneously trains CNNs and Swin-Transformers, enhancing their cross-supervised performance through the application of consistency loss. Gui et al. (2025) combined CNN and Transformer, the use of depth-separable convolution in the encoder and the introduction of a linear convolution module in the decoder enable efficient capture and fusion of multi-scale features. Zhong B. et al. (2025) proposed FERDNet, which combines the Multi-angle Feature Enhancement Module (MFE) and the High-Low Level Feature Information Compensation Module (ICM) to enhance the model's ability to extract road features. The above method combines local and global considerations of contextual features and detail features, thereby performing excellently in terms of intersection ratio, F1 score, and connectivity retention.

Coupling local and global road extraction networks harnesses the complementary strengths of CNNs and Transformers enabling comprehensive multi-scale feature extraction. While Transformer-based models have demonstrated superior global contextual reasoning, integrating them into such networks entails substantial drawbacks: excessive GPU memory consumption and quadratic computational complexity, which hinder real-time inference in resource-constrained scenarios. Furthermore, existing architectures adopt parallel branching strategies for global-local feature aggregation, merging contextual and fine-grained information without a principled mechanism guided by hierarchical feature importance and positional sensitivity. This indiscriminate fusion lacks adaptive integration logic, this leads to exponential growth in parameters due to redundant cross-feature interactions, as well as compromised computational efficiency due to unnecessary information mixing.

2.3 Lightweight design

In the quest for segmentation precision, the majority of road extraction networks have inadvertently led to an increase in model complexity, resulting in a substantial rise in network parameters. This surge impacts the model's efficiency and its practical deployability. Wei and colleagues Wei et al. (2020) introduced a framework that concurrently extracts both road surfaces and centerlines, utilizing an FCN for initial segmentation and further refining details with iterative lightweight FCN applications. Wang et al. (2024) enhanced segmentation accuracy by focusing on feature extraction through context fusion and self-learning sampling, effectively reducing redundancy and model complexity via dual feature fusion. Xiao et al. (2022) advocate for the RATT-UNet in mining road extraction, incorporating a RATT module that integrates residual connections with attention mechanisms to decrease the parameter count. Sun et al. (2022) tackle the issue of excessive parameters by introducing the LRSR-net, which leverages an extended joint convolution module to offset pooling layer losses and trim down the parameter list. Sultonov et al. (2022) crafted two lightweight networks for extracting road networks from drone imagery, melding UNet with depthwise separable convolutions, ConvMixer layers, and an initialization module. Han et al. (2023) proposed the target-aware LOANet, utilizing a lightweight dense connected network in its encoder. Zhao et al. (2025) designed S-MobileNet, combining 3D convolution, time series models, LSTM, and attention pooling mechanisms to extract and aggregate individual and group behavior characteristics. The LMSFFNet, presented in Yi et al. (2023), strikes a balance between execution speed and segmentation precision with a MobileViT backbone and a lightweight multi-scale context fusion module, thereby reducing parameter count and bolstering feature extraction capabilities. Liu et al. (2023) introduced LRDNet, a novel lightweight road detection method that employs a Multi-Scale Convolutional Attention Network (MSCAN) and a coupled decoder head to enhance detection speed and mitigate issues like occlusion and edge artifacts through expansive receptive field feature extraction and parallel decoding.

Lightweight road extraction networks have demonstrated the capability to enhance segmentation accuracy while reducing architectural complexity and parameter overhead, thereby boosting model efficiency and practical applicability in real-world scenarios. However, the pursuit of architectural parsimony create a critical challenge: existing methods struggle to balance model lightness with superior segmentation performance, as they typically sacrifice feature discriminability for computational economy. Maximizing parameter utilization efficiency to enable lightweight networks to simultaneously preserve global continuity and local details in remote sensing road extraction remains a critical challenge.

3 Methods

In this section, we will provide a detailed description of the construction of the Hierarchical Perception Lightweight Network for Road Extraction.

3.1 Overview of the network architecture

HPLNet employs a hierarchical perception strategy to effectively capture both global extensibility and local connectivity characteristics of roads. The global extensibility, which reflects

the semantic-level representation of road networks, is preserved in deep-level features, while the local connectivity, corresponding to detailed road structures, is maintained in shallow-level features. Based on this, we decompose the road input into hierarchical representations composed of shallow-level pixel features and deep-level semantic features, with the output features being the coupled shallow-level and deep-level road features, as shown in Equation 1.

$$\begin{cases} Out = g\left(f\left(In\right)\right) \\ f\left(In\right) = \left(Fea_S, Fea_D\right) \end{cases} \tag{1}$$

where Out represents the output, In represent the Input, $g(\cdot)$ represents a function that couples from deep and shallow features to obtain an output, $f(\cdot)$ represents the function that acquires deep and shallow features throughout the encoding and decoding process, Fea_S and Fea_D represent the shallow pixel features and deep semantic features.

Fea_S and Fea_D are generated by the function $f(\cdot)$ during the encoding and decoding processes, Fea_S includes shallow encoding features $Fea_S^{e,i}$ and shallow decoding features $Fea_S^{d,i}$, while Fea_D includes deep encoding features $Fea_D^{e,i}$ and deep decoding features $Fea_D^{d,i}$. The decoupling and coupling strategies for Fea_S and Fea_D are crucial to effectively leveraging both types of information for road extraction. We propose that employing distinct extraction strategies tailored to each type of information is necessary, followed by a fusion process that effectively integrates these multi-level representations.

To achieve comprehensive extraction of both Fea_S and Fea_D, we propose a hierarchical feature learning framework with specialized extraction parts: The Shallow Perception Part employs convolution-focused operations to capture local road characteristics, complemented by a channel-shift mechanism to enhance feature diversity. This branch specializes in learning low-level visual patterns (e.g., edges, textures, and corners) that are crucial for precise pixel-level road localization, thereby generating detailed segmentation outputs; The Deep Perception Part utilizes attention-focused mechanisms to model global road topology, similarly integrated with channel-shift operations. Positioned in deeper network layers, this branch processes high-level semantic information to understand complex road structures and connectivity patterns. As shown in Figure 2, the hierarchical architecture performs each corresponding encoding and decoding operation n times in the shallow perception part and deep perception part. In this study, we set $n = 2$ to systematically decouple and recouple road features in the network structure, thereby gradually refining local details and global context.

The encoding process for both shallow and deep features is defined in Equation 2,

$$\begin{cases} \mathrm{Fea}_S^{e,i} = \begin{cases} E_s\left(\mathrm{Initial\_conv}\left(In\right); \theta_S^{e,i}\right) & i = 1 \\ E_s\left(\mathrm{Fea}_S^{e,i}; \theta_S^{e,i+1}\right) & 1 < i \le n \end{cases} \\ \mathrm{Fea}_D^{e,i} = \begin{cases} E_d\left(\mathrm{Fea}_S^{e,i}; \theta_D^{e,i}\right) & i = 1 \\ E_d\left(\mathrm{Fea}_D^{e,i}; \theta_D^{e,i+1}\right) & 1 < i \le n \end{cases} \end{cases} \tag{2}$$

where $Fea_S^{e,i}$ represent shallow encoding features, $Fea_D^{e,i}$ represent deep encoding features, E_s represent shallow encoding operation (3× three conv and channel shift), E_d represent deep encoding operation (ASCA and channel shift), $Initial\_conv(\cdot)$ represent initial convolution on input, In represent input image, $\theta_S^{e,i}, \theta_D^{e,i}$ represent parameters for shallow and deep encoding.

The decoding process is defined in Equation 3,

$$\begin{cases} \mathrm{Fea}_D^{d,i} = \begin{cases} D_d\left(\mathrm{Fea}_D^{e,i}; \theta_D^{d,i}\right) & i = n \\ D_d\left(\mathrm{Fea}_D^{d,i} + \mathrm{Fea}_D^{e,i}; \theta_D^{d,i-1}\right) & 1 < i \le n \end{cases} \\ \mathrm{Fea}_S^{d,i} = \begin{cases} D_s\left(\mathrm{Fea}_D^{d,i} + \mathrm{Fea}_S^{e,i}; \theta_S^{d,i}\right) & i = n \\ D_s\left(\mathrm{Fea}_S^{d,i} + \mathrm{Fea}_S^{e,i}; \theta_S^{d,i-1}\right) & 1 < i \le n \end{cases} \end{cases} \tag{3}$$

where $\mathrm{Fea}_D^{d,i}$ represent deep decoding features, $\mathrm{Fea}_S^{d,i}$ represent shallow decoding features, D_d represent deep decoding operation (ASCA and channel shift), D_s represent shallow decoding operation (strip conv and channel shift), $\mathrm{Fea}_D^{e,i}$ represent deep encoding features, $\mathrm{Fea}_S^{e,i}$ represent shallow encoding features, $\theta_D^{d,i}, \theta_S^{d,i}$ represent parameters for deep and shallow decoding.

Road segmentation results (out) is shown in Equation 4,

$$Out = Final\_conv\left(Fea_S^{d,i-1}\right) \tag{4}$$

where Out represents the output and $Final\_conv$ is the function that obtains the final output by convolution and sigmoid.

3.2 Adaptive spatial-channel attention module (ASCA)

Road extraction demands precise modeling of both local details and global structures to capture the inherent continuity of road networks. Self attention mechanisms effectively model such contextual dependencies via pairwise spatial correlations but incur prohibitive quadratic complexity, rendering them impractical for large scale road imagery. For real time applications, lightweight design is imperative to balance efficiency and performance. To overcome this bottleneck, we propose a Lightweight Adaptive Spatial-Channel Attention module, building on the dual aggregation framework in Chen et al. (2023) with targeted efficiency optimizations. We replace the original alternating dual-aggregation Transformer blocks with a sequential spatial channel attention architecture, significantly reducing computational overhead without sacrificing representational capacity. The ASCA module operates in two sequential stages to model spatial and channel dependencies, enabling efficient capture of both local details and global context for comprehensive road feature extraction.

Specifically, as shown in Figure 3a, the features are first linearly projected via convolution and one branch reshapes the features into the feature maps $Y_s \in \mathbb{R}^{h\times HW\times C}$, which is decomposed to as the spatial-based vectors,i.e., Query ($Q_s \in \mathbb{R}^{h\times HW\times C}$), Key ($K_s \in \mathbb{R}^{h\times HW\times C}$) and Value ($V_s \in \mathbb{R}^{h\times C\times HW}$) matrices, where h represents the number of attention heads and C represents the number of feature channels. The spatial self-attention mechanism is then applied to extract spatial information from the image to obtain the spatial-wise self-attention feature maps Y_s, with the spatial self-attention mechanism defined as Equation 5,

$$Y_s = \mathrm{softmax}\left(\frac{Q_s K_s^{\mathrm{T}}}{\alpha}\right) V_s, \tag{5}$$

where the learnable scaling parameter α is used to modulate the inner product before applying the softmax function.

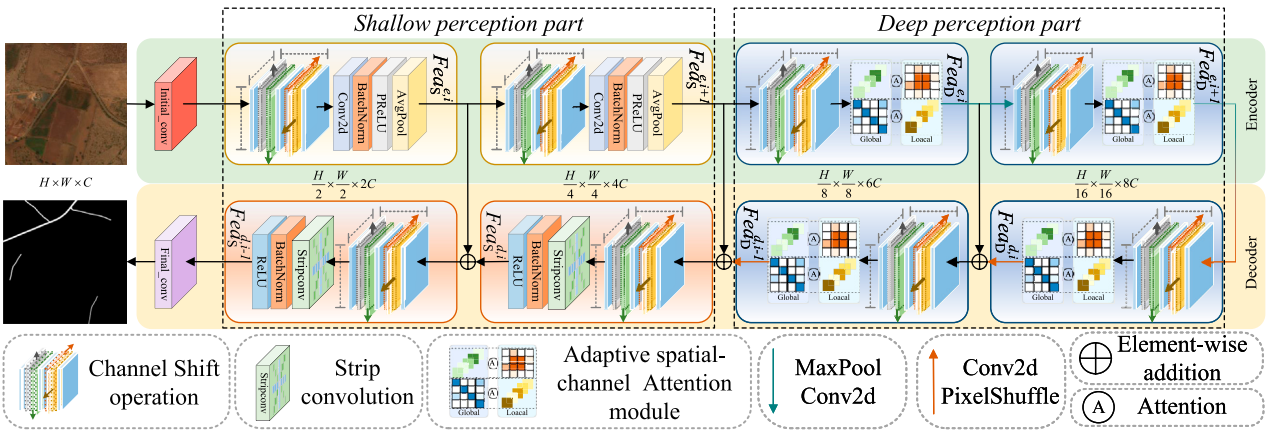

FIGURE 2
Overall architecture of the proposed HPLNet. The network employs shallow perception to extract and upscale local features for fine-grained segmentation and deep perception with an adaptive spatial-channel attention mechanism to enhance global feature representation and segmentation accuracy.

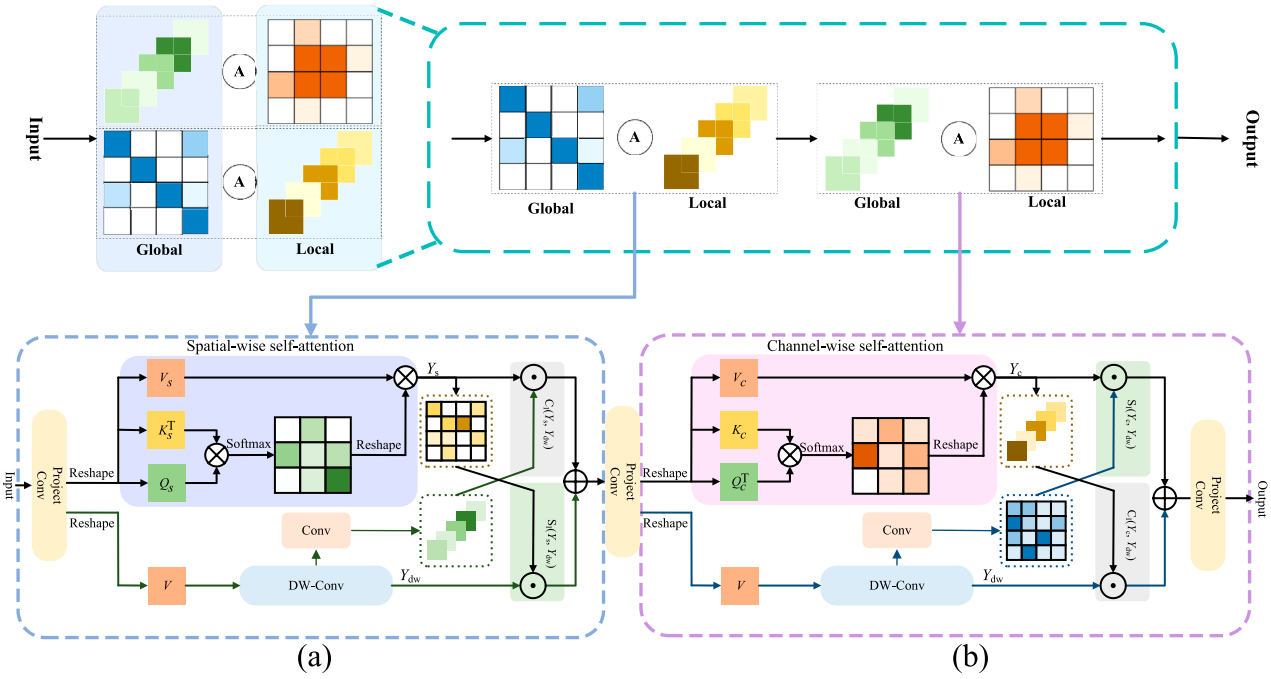

FIGURE 3
An overview of Adaptive spatial-channel attention module. The module is a lightweight design that sequentially integrates a dual-aggregation attention interaction mechanism to extract global features. Outs denotes the output result of **(a)** and Outc denotes the output result of **(b)**.

Unlike the global feature information captured by the previous branch, the other branch uses depth-wise separable convolution to extract local feature. To compensate for the dimensionality singularity of the information attended to by the two branches, we perform channel-wise interaction and spatial-wise interaction on the two types of perceived information. The spatial and channel weights are adaptively integrated into the two branches and after pixel-wise addition and projection convolution, the final output is obtained.

In Figure 3b, similar operations are performed. One branch reshapes the features into $Y_c \in \mathbb{R}^{h\times HW\times C}$, which is decomposed to as the Query (Q_s^T), Key (K_s) and Value (V_c) matrices. The channel self-attention mechanism is then used to extract channel information from the image, with the channel self-attention mechanism defined as Equation 6:

$$Y_c^T = \text{softmax}\left(\frac{Q_c^T K_c}{\beta}\right)V_c^T, \tag{6}$$

where the learnable scaling parameter β is used to modulate the inner product before applying the softmax function.

To address the dimensionality singularity of the information attended to by both branches, ASCA is able to aggregate spatial and channel features by means of (a) and (b) cascades, resulting in a robust feature representation. Figure 3a illustrates the modeling of the distant spatial context, which strengthens the spatial representation within each feature map. Figure 3b, on the other hand, demonstrates an improved construction of channel dependencies. The modeling of the global channel context shown in Figure 3b reciprocally aids in capturing spatial features and expanding the receptive field as depicted in Figure 3A. The channel interactions and spatial interactions are described by Equation 7:

$$\begin{aligned} S_I(M,N) &= M \odot Map_{\mathbf{s}}(N), \\ C_I(M,N) &= M \odot Map_{\mathbf{c}}(N), \end{aligned} \tag{7}$$

where C_I denotes channel interactions and S_I denotes spatial interactions. M, N are the input features, $M, N \in \mathbb{R}^{H\times W\times C}$, $\odot$ denotes elemental-wise multiplication, and $Map_{\mathbf{s}}$ and $Map_{\mathbf{c}}$ denote the spatial and channel feature maps of N, respectively.

The spatial and channel weights are adaptively integrated into both branches and after pixel-wise addition and projection convolution, the final output is obtained. This process can be described by the following Equation 8:

$$\begin{aligned} Out_s(In) &= (C_I(Y_{\mathbf{s}}, Y_{\mathbf{dw}}) + S_I(Y_{\mathbf{dw}}, Y_{\mathbf{s}}))W_p, \\ Out_{\mathbf{c}}(Out_{\mathbf{s}}) &= (S_I(Y_{\mathbf{c}}, Y_{\mathbf{dw}}) + C_I(Y_{\mathbf{dw}}, Y_{\mathbf{c}}))W_p, \end{aligned} \tag{8}$$

where In denotes the input feature, $Y_{\mathbf{s}}$ denotes the feature that has undergone spatial self-attention, $Y_{\mathbf{c}}$ denotes the feature that has undergone channel self-attention, $Y_{\mathbf{dw}}$ is the feature that has undergone depth-separable convolution, and W_p is the linear projection that is used to fuse all the features, $Out_{\mathbf{s}}$ denotes the output result of Figure 3a and $Out_{\mathbf{c}}$ denotes the output result of Figure 3b.

The computational complexity of self-attention mechanism is positively correlated with the square of the number of pixel patches, we have taken compensatory measures in this module so that we can greatly reduce the vector dimensions and the number of parameters in the self-attention mechanism, thus achieving a lightweight design. In conclusion, through the integration of a lightweight dual-aggregation attention mechanism, we facilitate the mapping of global and local perception information interactions across both spatial and channel dimensions. This approach effectively extracts comprehensive global features of roads, enhancing the overall understanding and representation of the data.

3.3 Channel shift operation module

In response to the prevalent challenges of road disconnection and complex orientations in remote sensing imagery, we introduce a lightweight and parameter-free module: the Channel Shift Operation Module Zhang et al. (2023a). The core idea is to explicitly enlarge the effective receptive field of a neuron by spatially shifting its corresponding feature channels. This mechanism enhances the model's ability to perceive local contextual details by creating a set of spatially variant feature representations. It encourages the subsequent attention module to operate on a richer feature space, where subtle structural and textural cues from neighboring regions are explicitly incorporated, thereby strengthening the feature discrimination power without introducing any learnable parameters or additional computational burden. The channel shift operation is shown in Figure 4, the operation is defined as Equation 9:

$$\Pi = \left\{\pi_{\mathrm{U}}^{d}, \pi_{\mathrm{D}}^{d}, \pi_{\mathrm{L}}^{d}, \pi_{\mathrm{R}}^{d}\right\}, d < T_m, \tag{9}$$

where Π represents the specific operation function, T_m is a preset threshold that limits the degree of channel shift to avoid loss of spatial locality in the image. π_i^d denotes the channel shift operation (i indicates four directions of channel shifting and d indicates the magnitude of shift in a certain direction).

In addition, considering the lightweight design, we group the channels before performing the inter-group alignment, and the shifted channel groups are those close to the middle position. $\hat{\mathbf{s}}$ indicates the channel characteristics after performing the channel shift operation, refer to Equation 10:

$$\hat{\mathbf{s}} = \{\overbrace{0,\ldots}^{\pi_O^0}, \overbrace{S_i,\ldots}^{\pi_U^d}, \overbrace{S_{i+c+1},\ldots}^{\pi_D^d}, \overbrace{S_{i+2c+1},\ldots}^{\pi_L^d}, \overbrace{S_{i+3c+1},\ldots}^{\pi_R^d}, \overbrace{S_{i+4c+1},\ldots}^{\pi_O^0},\} \tag{10}$$
$$\underbrace{}_{\frac{(C-4c)}{2}} \quad \underbrace{}_{c} \quad \underbrace{\phantom{S_{i+c+1},\ldots}}_{c} \quad \underbrace{\phantom{S_{i+2c+1},\ldots}}_{c} \quad \underbrace{\phantom{S_{i+3c+1},\ldots}}_{c} \quad \underbrace{\phantom{S_{i+4c+1},\ldots}}_{\frac{(C-4c)}{2}}$$

where π_O^0 indicates no channel shift operation or a shift magnitude of zero, π_i^d indicates that the channel is displaced in four directions: up, down, left, and right, with a magnitude of d, $\frac{(C-4c)}{2}$ and c indicates the number of channels with or without channel shift operation.

To investigate the impact of varying shift channels and shift pixels, we conducted a series of experiments, the results of which are summarized in Figure 5. Specifically, under a consistent training protocol, we systematically compared the training loss and intersection over union (IoU) metrics. Our parameter selection strategy involved identifying the configuration that simultaneously minimized the training loss and maximized the IoU, thereby optimizing the model's performance. In the end, we set the number of shift channels is two and the amplitude d of the channel shift to two pixel in order to capture this difference.

3.4 Strip convolution module

The inherent linearity of road networks demands a large receptive field for accurate representation. Standard convolutional kernels, however, are square and thus geometrically incongruent with road structures. This misalignment introduces significant noise from irrelevant background pixels, which can degrade feature quality. To address this limitation, we introduce strip convolutions Sun et al. (2019), kernels designed to be geometrically congruent with the linear geometry of roads. By focusing exclusively on the linear structures, strip convolutions effectively filter out peripheral noise, thereby improving the precision and robustness of road feature extraction.

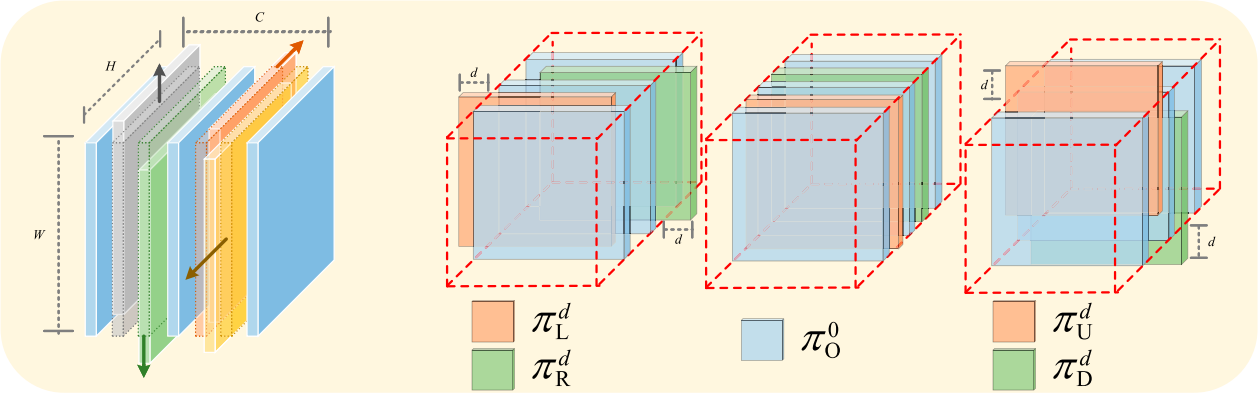

FIGURE 4
The channel shift operation is shown in the figure. π_{O}^{0} represents the original feature without movement, π_{L}^{d}, π_{R}^{d}, π_{U}^{d} and π_{D}^{d} represent the features obtained after the channel is shifted left, right, up and down, respectively. d indicates the magnitude of shift in a certain direction.

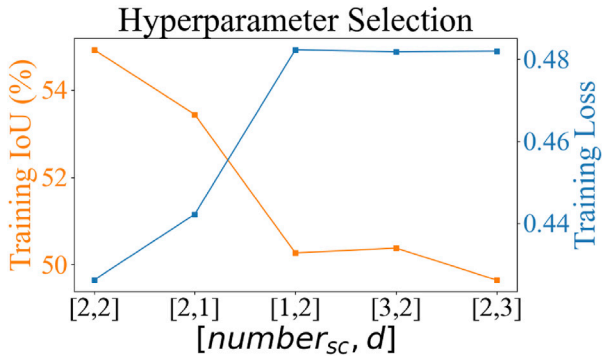

FIGURE 5
Hyperparameter selection chart, left vertical axis [Training IoU], right vertical axis [Training Loss], Horizontal axis ($number_{sc}$, d), where $number_{sc}$ represents the number of channels shift, and d represents the magnitude of the channel shift.

Specifically, let $k \in \mathbb{R}^{2r+1}$ denotes the 1D convolution filter of size $2r+1$ and $y_I \in \mathbb{R}^{H\times W}$ be the result of 1D transpose convolution of input $x \in \mathbb{R}^{H\times W}$ and the filter k at direction $I = (I_{\mathrm{h}}, I_{\mathrm{w}})$. We have

$$y_I[i,j] = (x \otimes k)_I = \sum_{t=-r}^{r} x[i + I_{\mathrm{h}}t, j + I_{\mathrm{w}}t] \cdot k[r-t], \tag{11}$$

where $\otimes$ is the convolution operation and I is the direction indicator vector of the 1D filter, which takes four values $(0,1)$, $(1,0)$, $(1,1)$, $(-1,1)$ for horizontal, vertical, forward diagonal and backward diagonal transpose convolution, respectively, shown in Figure 6 and Equation 11.

3.5 Loss function

In this work, we use the BceDiceLoss Mei et al. (2021), which synergistically integrates the merits of BCE Loss and Dice Loss. The loss function is shown as Equation 12:

$$\begin{aligned} L_{\mathrm{BceDice}} &= L_{\mathrm{BCE}} + L_{\mathrm{Dice}} \\ L_{\mathrm{BCE}} &= -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)] \\ L_{\mathrm{Dice}} &= 1 - \frac{2|A \cap B|}{|A| + |B|} \end{aligned} \tag{12}$$

where N is the total number of samples, y_i is the true label of the i-th sample, and $\hat{y}_i$ is the predicted probability of the i-th sample being positive, A represents the predicted segmentation region, B represents the true segmentation region, $|A \cap B|$ is the area (or pixel count) of the intersection of A and B, and $|A|$ and $|B|$ are the areas (or pixel counts) of A and B, respectively.

This composite loss function effectively attends to both the pixel prediction error and the region similarity. As a result, it significantly enhances the performance of the model in image segmentation tasks, enabling more accurate and robust segmentation results.

4 Experiments

In this section, we present the datasets, evaluation metrics and experimental setup. Subsequently, we assess our method on remote sensing datasets. Furthermore, we conduct ablation studies to validate our key innovations.

4.1 Datasets

We chose to validate the effect of our proposed HPLNet in three datasets. The three datasets are described below:

DeepGlobe Demir et al. (2018): This dataset consists of high-resolution images with a spatial resolution of 50 cm/pixel, each image having a size of 1024 × 1024 pixels. The images cover three regions: Thailand, Indonesia and India. The dataset provides detailed pixel-level annotations, distinguishing between road and background classes. The dataset consists a total of 6,226 images are included. Following the methodology in Mei et al. (2021), we divided these images into a training set with 4,696 images and a test set with 1,530 images.

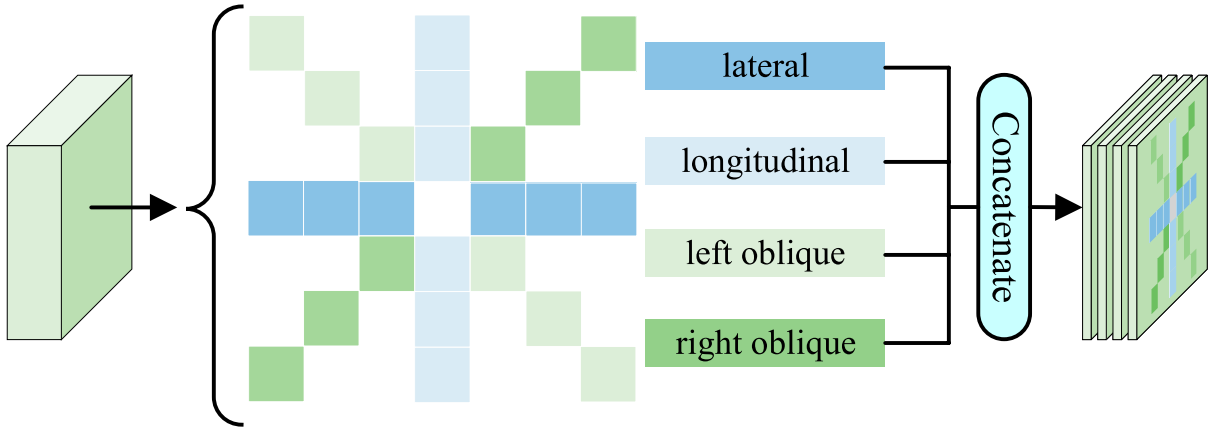

FIGURE 6
The strip convolution block. This block encompasses four distinct orientations: lateral, longitudinal, left oblique and right oblique.

Massachusetts Mnih (2013): The Massachusetts Road Dataset contains aerial imagery of Massachusetts with a resolution of 1 m/pixel. It includes 1,108 training images, 49 test images and 14 validation images. Each image has a resolution of 1500 × 1500 pixels.

CHN6-CUG Zhu et al. (2021): The CHN6-CUG dataset selected six representative cities in China and marked roads including railways, highways, urban roads, and rural roads. It contains 4,511 marked images with a size of 512 × 512, divided into 3,608 images for model training and 903 images for testing and result evaluation, with a resolution of 50 cm/pixel.

4.2 Evaluation metrics

To comprehensively evaluate the model's lightweight design and segmentation performance, we assess the model size and complexity using Param (parameter count) and FLOPs (floating-point operations). We also evaluate performance using widely accepted metrics, such as Intersection over Union (IoU) and F1-score.

- Param (Parameter Count): This refers to the total number of learnable parameters in the model, providing an indication of the model's size and complexity.
- FLOPs (Floating-Point Operations): This measures the number of floating-point operations required to perform the model's calculations, reflecting the computational complexity and operational efficiency of the model.
- Inference time: Inference time refers to the time required for the method to complete calculations and generate output results after receiving input data.
- Intersection over Union (IoU): This metric measures the overlap between predicted segmentation and true labels in image segmentation tasks.
- F1-Score: The harmonic mean of precision and recall, used to evaluate the performance of binary classification models.

4.3 Experimental setup

Our proposed model was implemented using PyTorch 1.13 on an NVIDIA 4090 with 24 GB of memory. The training batch size was set to 22, the optimizer used was AdamW, the initial learning rate was set to 0.001 and the weight decay coefficient was set to 0.01. We did not perform any preprocessing operations such as image normalisation, scaling, or class balancing, nor did we perform any cropping. The network is adaptive to the size of the image input. The input and output sizes of Deepglobe are 1024 pixels × 1024 pixels. The input and output sizes of Massachusetts are 1500 pixels × 1500 pixels and 1024 pixels × 1024 pixels, respectively. Additionally, we did not perform data augmentation operations such as rotation, flipping, scaling, or brightness adjustment.

4.4 Comparison with mainstream models

4.4.1 Comparison on public datasets

To comprehensively evaluate the segmentation performance of our model on the DeepGlobe and Massachusetts datasets, we have selected several visual results from both the compared models and our proposed method. The models included in the comparison are FCN Long et al. (2015), U-net Ronneberger et al. (2015b), DlinkNet Zhou et al. (2018), CoANet Mei et al. (2021), RcfsNet Yang Z. et al. (2023), SDsegNet Lin et al. (2024), VisionMamba-Unet-2 (VMUnet-2)Zhang et al. (2024), DeNet Guo et al. (2025), CSWin-UNet Liu et al. (2025) and PMFSNet Zhong J. et al. (2025).

4.4.2 Qualitative comparison in DeepGlobe dataset

As shown in Figure 7, the segmentation results of FCN and U-net for both cyan and black roads clearly demonstrate the subpar performance of these two methods. Dlinknet also has connectivity issues in the partitioning results. CoANet exhibits an over-segmentation issue in the segmented images. RcfsNet suffers from mis-segmentation in the first image, and in the second and fourth images, it fails to segment long-distance road regions. SDSeg

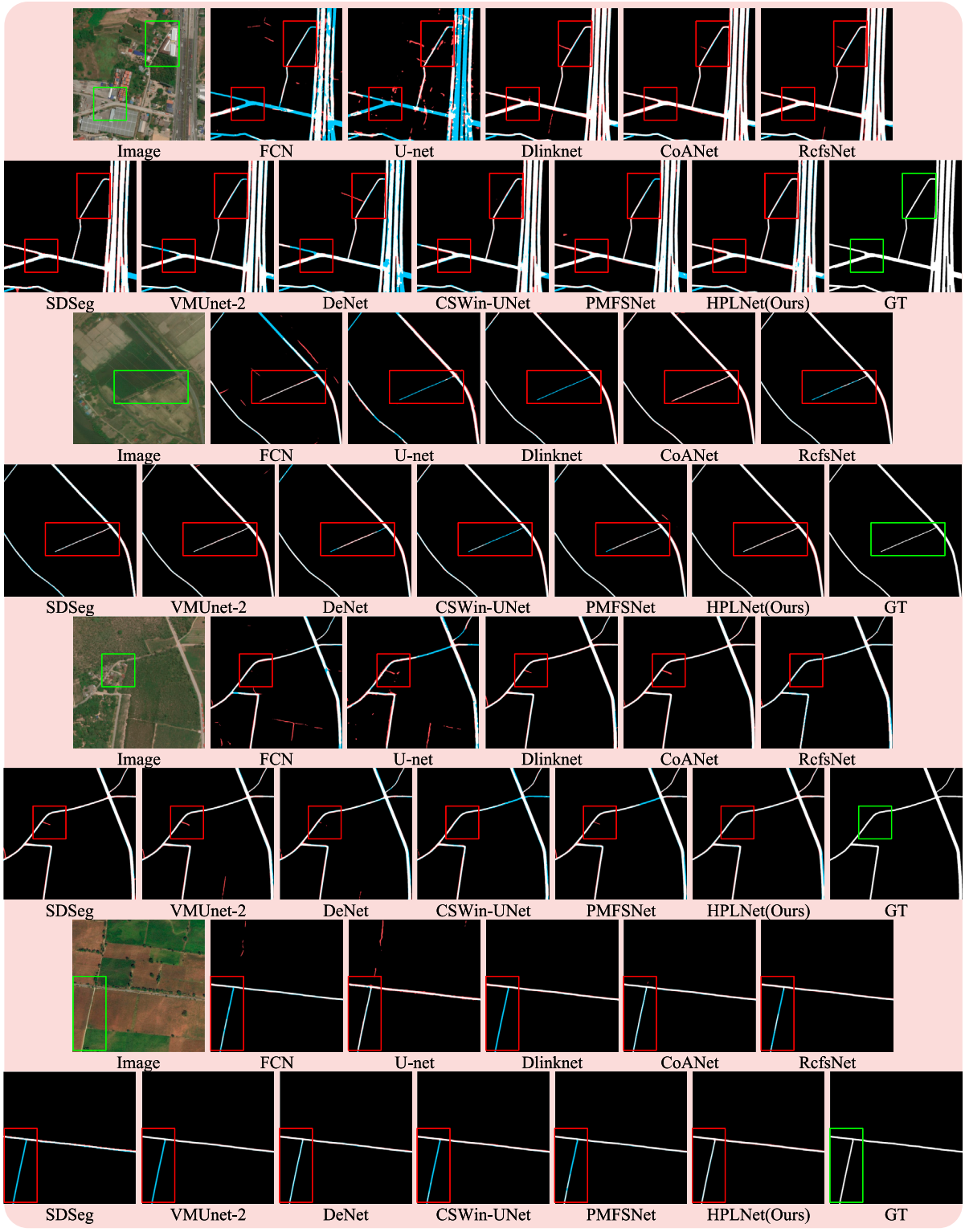

FIGURE 7
Qualitative comparison with other state-of-the-art methods in DeepGlobe dataset. White: true positive, black: false positive, cyan: false negative.

produces incorrect segmentations in the third image and misses segments in the fourth image. VMUnet-2 shows missed segmentations in the first and fourth images, along with incorrect segmentations in the third image. DeNet retained the edges in the segmentation results, but there were still some minor errors and disconnections in connectivity in the fine details. PMFSNet still has some issues with segmentation and connection breaks in the details. In contrast, HPLNet yield more accurate and visually appealing

TABLE 1 Quantitative comparison of our proposed HPLNet with some advanced road extraction methods on DeepGlobe dataset. Numbers in bold indicate the best values, while numbers with the underline indicate the second good values in each column.

Methods	IoU (%)	F1-score (%)
FCN	44.12	53.63
U-net	46.10	63.37
DlinkNet	64.37	77.16
CoANet	63.31	78.72
RcfsNet	**67.40**	79.64
SDseg	60.07	75.05
VMUnet-2	65.62	79.24
DeNet	63.04	77.33
CSWin-Unet	64.92	78.42
PMFSNet	65.06	78.83
HPLNet (Ours)	66.61	**79.96**

segmentation results, outperforming the other methods in terms of overall segmentation quality.

4.4.3 Quantitative comparison in DeepGlobe dataset

As shown in Table 1, the IoU values and F1 scores shown in the table agree well with the results of our qualitative analysis. On the Deepglobe dataset, FCN and U-net perform poorly in terms of IoU and F1-score, and the rest of the compared methods have IoU values above 0.60, with HPLNet lagging behind RcfsNet by 0.0079, which may be attributed to the fact that RcfsNet utilises multiscale contextual extraction and the full-stage feature fusion module with a significant increase in the parameters obtains better segmentation results; In the results of CSWin-Unet, there are some issues with incomplete segmentation. Compared to the sliding window attention mechanism, the ASCA we introduced is better suited to the task of road segmentation. Notably, HPLNet attains the highest F1 score, surpassing the method ranking second, RcfsNet, by 0.0032. This indicates that our approach strikes an optimal balance between accuracy and comprehensiveness, thereby demonstrating greater reliability in semantic segmentation tasks.

4.4.4 Qualitative comparison in Massachusetts dataset

As shown in Figure 8, in the upper segmented image, the FCN produces numerous cyan roads in its segmentation result, exhibiting a pronounced under-segmentation issue. Meanwhile, the segmentation outcomes of Dlinknet, CoANet, RcfsNet, DeNet and PMFSNet display black roads in the central region of the image, indicating severe over-segmentation. Notably, only our proposed method can accurately segment the curved paths within the black bounding box. For the lower image, FCN continues to suffer from under-segmentation, while the segmentation result of U-Net contains an abundance of extraneous pixels. Within the black bounding box, CoANet, RcfsNet, VMUnet-2, and PMFSNet fail to successfully segment the upward road. In contrast, both Dlinknet and our HPLNet manage to segment the road. A comprehensive comparison of the segmentation results between Dlinknet and HPLNet reveals that HPLNet captures finer details, especially in the bottom-right corner of the image.

4.4.5 Quantitative comparison in massachusetts dataset

As shown in Table 2, the IoU values and F1-score presented in the table align closely with our qualitative analysis. Specifically, FCN exhibits the lowest IoU and F1-score values, while our HPLNet achieves the highest F1-score and IoU on the Massachusetts dataset. This quantitative evidence validates the superior visual segmentation quality demonstrated by HPLNet. Compared with DeNet, HPLNet achieves improvements in IoU and F1-score, compared with RcfsNet, HPLNet achieves improvements of 0.0121 in IoU and 0.0091 in F1-score, which vividly showcase its preeminent performance in road segmentation. These quantitative results not only affirm the reliability of our proposed method in precisely matching road segmentation outputs with ground - truth labels on the Massachusetts dataset but also strongly emphasize its effectiveness in handling semantic segmentation tasks.

4.4.6 Model size and complexity evaluation

To evaluate the efficiency of our proposed method, we conducted a comparative analysis with several state-of-the-art models, focusing on key metrics including parameter count, floating-point operations (FLOPs), and inference latency. All experiments were performed under a unified input setting of 1024 × 1024 images, with detailed results summarized in Table 3. The models included in the comparison are FCN, U-net, DlinkNet, CoANet, RcfsNet, SDsegNet, VisionMamba-Unet-2 (VMUnet-2), DeNet, CSWin-UNet and PMFSNet.

As seen in Table 3, SDSeg, utilizing a diffusion model, has the highest FLOPs and the slowest inference time among the methods considered. RcfsNet, which uses multi-scale context extraction and full-stage feature fusion, has the largest number of parameters. VMUnet-2, based on the Mamba framework, has the lowest FLOPs, benefiting from the efficiency of its state-space model. Although RcfsNet achieves the highest segmentation accuracy, it does so at a high computational cost, requiring 76.74M parameters and 182.36G FLOPs. PMFSNet achieves the lowest FLOPs and the fastest inference speed by simplifying the computational complexity of the hierarchical structure based on unet and the self-attention mechanism, but this operation will lead to a decline in the segmentation effect. However, our method is designed to significantly reduce the number of parameters and FLOPs, accelerated inference time, while only slightly decreasing the IoU on the DeepGlobe dataset. Taking into account the complexity of the method and the segmentation accuracy, HPLNet effectively combines high segmentation accuracy with a lightweight design. It offers the benefits of a compact model with superior accuracy, making it particularly suitable for mainstream remote sensing image road extraction tasks.

4.5 Ablation study

We conducted an ablation study to assess the impact of each module on the size and complexity of the method in qualitative and

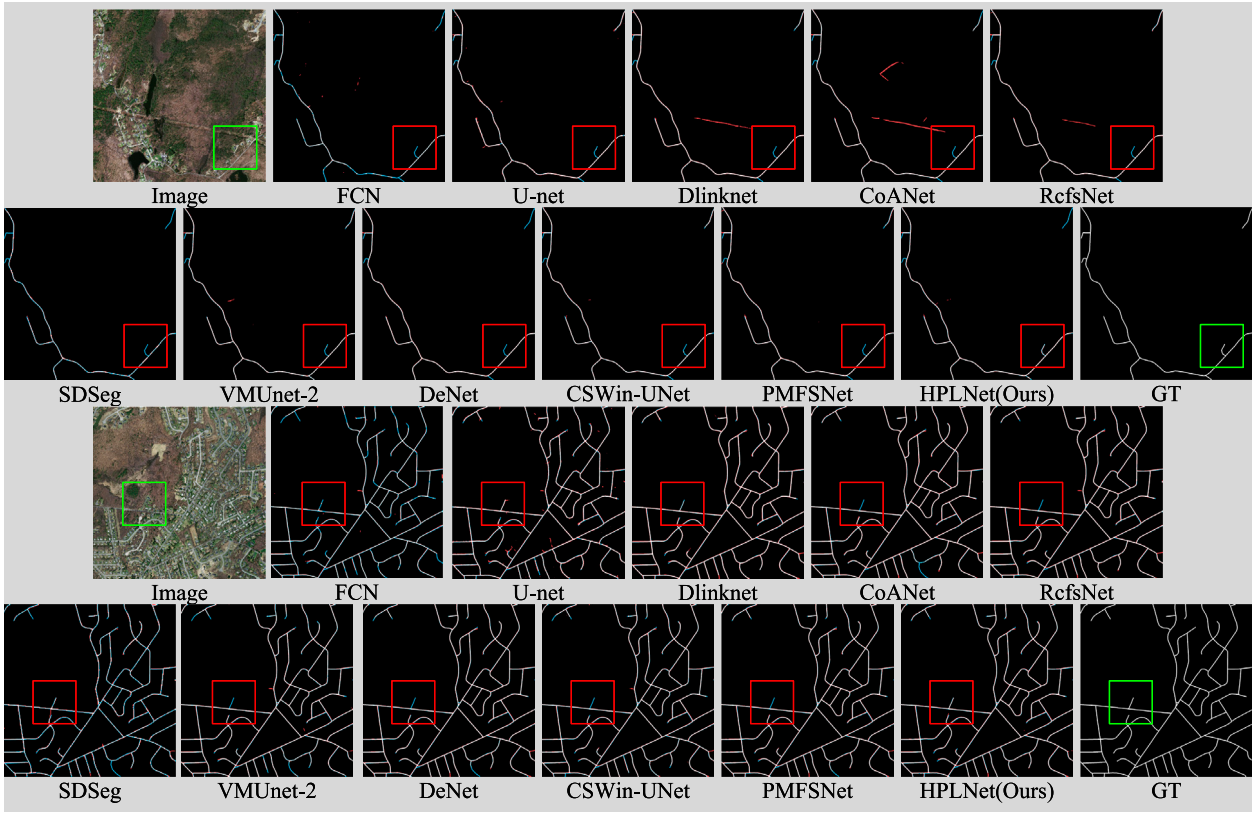

FIGURE 8
Qualitative comparison with other state-of-the-art methods in Massachusetts dataset. White: true positive, black: false positive, cyan: false negative.

TABLE 2 Quantitative comparison of our proposed HPLNet with some advanced road extraction methods on Massachusetts dataset. Numbers in bold indicate the best values, while numbers with the underline indicate the second good values in each column.

Methods	IoU (%)	F1-score (%)
FCN	39.62	56.75
U-net	51.28	67.80
DlinkNet	61.46	76.13
CoANet	61.86	76.44
RcfsNet	63.11	77.38
SDseg	51.05	67.59
VMUnet-2	58.12	73.51
DeNet	<u>64.19</u>	<u>78.19</u>
CSWin-Unet	63.50	74.89
PMFSNet	63.14	77.41
HPLNet (Ours)	**64.32**	**78.29**

TABLE 3 Quantitative comparison between our proposed HPLNet and some advanced extraction methods in terms of FLOPs, parameters and inference time. Numbers in bold indicate the best values, while numbers with the underline indicate the second good values in each column.

Methods	FLOPs(G)	Params(M)	Inference time(s)
FCN	101.96	18.64	0.0729
U-net	124.48	13.39	0.0944
DlinkNet	36.31	31.10	0.0940
CoANet	277.41	59.15	0.1192
RcfsNet	182.36	76.74	0.1860
SDseg	541.16	34.15	1.3176
VMUnet-2	**11.17**	14.40	0.0827
DeNet	159.09	25.52	0.1407
CSWin-Unet	98.68	23.57	0.0973
PMFSNet	46.44	**0.99**	**0.0517**
HPLNet (Ours)	<u>20.11</u>	<u>2.88</u>	<u>0.0602</u>

quantitative experiments on both the DeepGlobe and Massachusetts datasets. After introducing or not introducing ASCA, band convolution and channel shift operations, the obtained method is as follows: A, B, C, D, E, and F. Method-A is the baseline model and does not contain any modules; in method-B, we introduced the ASCA module. Method-C contains strip convolution with shallow decoding and ASCA; Method-D shows the replacement of the deep and shallow 3× three convolutions with strip convolution

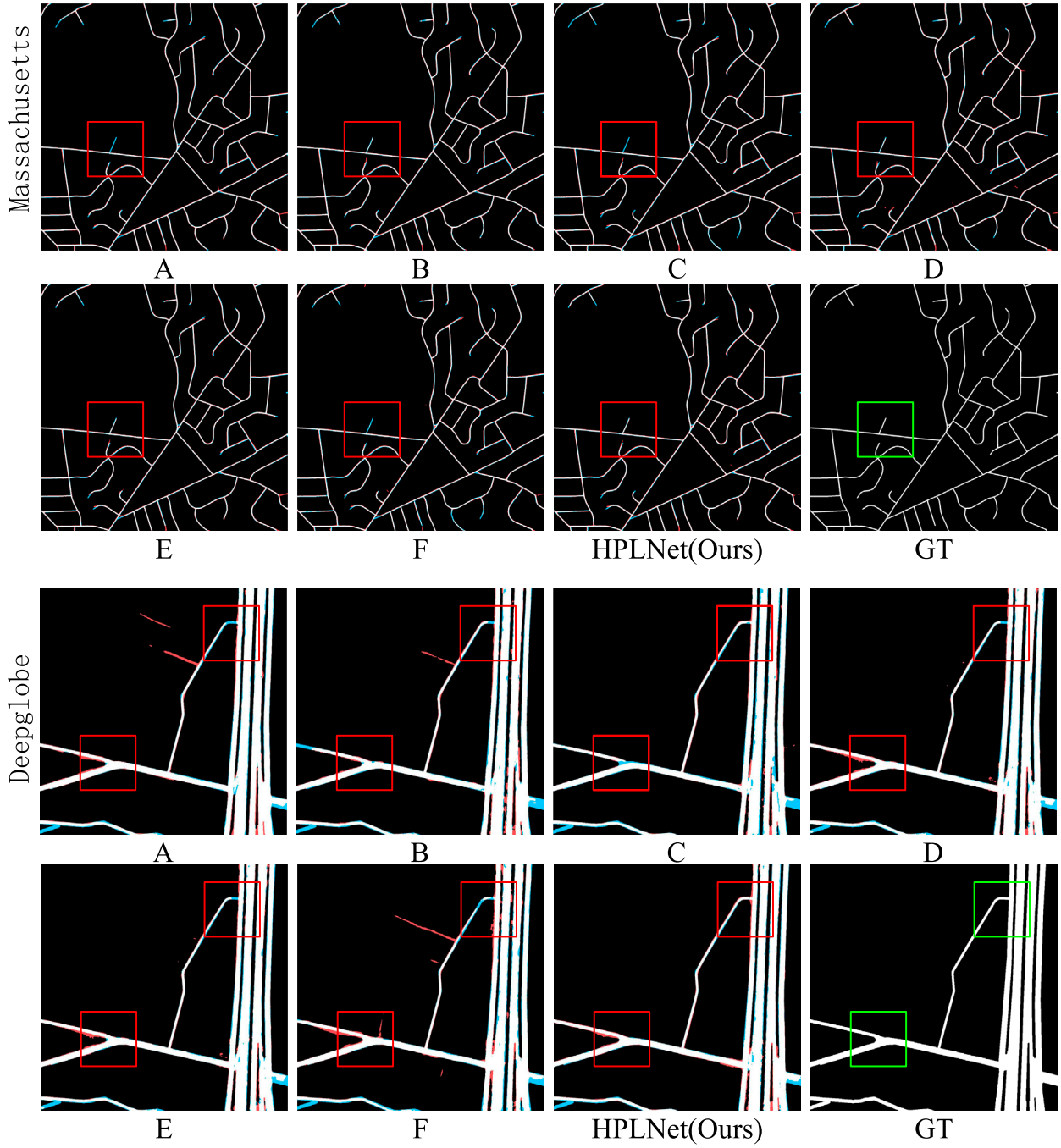

FIGURE 9
HPLNet visualisation results on DeepGlobe and Massachusetts datasets. **(A)** none, **(B)** ASCA, **(C)** strip convolution with ASCA, **(C)** strip convolution with ASCA. **(D)** strip convolution, **(E)** channel shift operation and ASCA, **(F)** channel shift and strip convolution. HPLNet (ours): HPLNet visualisation: strip convolution, channel shift operation and ASCA, GT: ground truth.

throughout the modelling process; Method-E contains the channel shift operation and ASCA; Method-G contains the channel shift operation and the strip convolution for shallow decoding.

Qualitative comparison is shown in Figure 9, the ASCA introduced in Method-B, Method-C and Method-E plays a positive role in preserving the global information of the images, including maintaining stronger road connectivity, reducing some false segmentations, and improving segmentation of finer road branches. Method-C builds on Method-B by using strip convolution instead of the 3×3 convolution in the shallow decoding part, further enhancing the model's ability to capture road connectivity and extension features. Method-D replaces the 3×3 convolution in the network architecture with the striped convolution, however, the introduction of striped convolution reduces the number of parameters but negatively affects the segmentation results, validating the advantages of our designed

TABLE 4 Quantitative results of ablation experiments, where CSO stands for Channel Shift operation module, Conv represents convolution, Strip represents strip convolution and ASCA denotes the Adaptive spatial-channel attention module. Numbers in bold indicate the best values, while numbers with the underline indicate the second good values in each column.

Methods	CSO	Conv	ASCA	Deepglobe		Massachusetts		FLOPs(G)	Params(M)
				IoU(%)	F1-score(%)	IoU(%)	F1-score(%)		
A	×	3 × 3[1]	×	64.44	78.37	62.88	77.21	4.55	0.63
B	×	3 × 3	√	64.85	78.68	63.13	77.40	5.41	3.01
C	×	*Strip*	√	<u>65.43</u>	<u>79.10</u>	<u>63.85</u>	<u>77.94</u>	5.04	2.88
D	×	*Strip*∗[2]	×	60.00	75.00	61.83	76.41	**2.80**	**0.10**
E	√	3 × 3	√	65.23	78.96	63.59	77.74	5.41	3.01
F	√	*Strip*	×	62.30	76.77	61.20	75.93	<u>4.25</u>	<u>0.51</u>
HPLNet (Ours)	√	*Strip*	√	**66.61**	**79.84**	**64.32**	**78.29**	5.10	2.88

[a]The use of 3 × 3 corresponds to the convolution method in shallow feature encoding.
[b]*Strip*∗ indicates that strip convolution is used instead of 3× three convolution for deep and shallow perception parts.

approach of introducing striped convolution only in the decoding part of shallow perception.

Quantitative comparison is shown in Table 4, indicate that the introduction of the ASCA increases the parameter count slightly, but the total number of parameters remains smaller than in previously compared models, demonstrating the lightweight nature of this attention mechanism. Meanwhile, the introduction of the channel shift operation further reduces the FLOPs and the number of parameters by adding zero padding after the shift to reduce the attention and convolution computation on the corresponding channel. In addition, the introduction of strip convolution leads to a reduction in the number of parameters and FLOPs. This achieves an optimal trade-off between model performance and efficiency.

In contrast to methods C and F, ASCA exerts a significantly more pronounced influence on segmentation performance. Furthermore, a comparison between methods B and E underscores the efficacy of the introduced channel shift operation in refining segmentation results, along with its ability to detect and rectify trailing branches.

5 Discussion

As shown in Figure 10, HPLNet demonstrates remarkable performance in edge segmentation. However, in a small number of images, lightweight attention perception strategy (ASCA) imposes constraints on maintaining global contextual connectivity, leading to subtle limitations in capturing long-range dependencies and structural consistency. Therefore, further improvements are needed to extract long-range dependencies from the road while maintaining a lightweight network design.

Furthermore, the generalization capability of a method is intrinsically linked to its practicality, reliability, and scalability. Consequently, we conducted generalization experiments to ascertain that the method's performance remains robust and does not degrade significantly when deployed in real-world scenarios characterized by potentially different data distributions. Specifically, we performed these evaluations on the CHN6-CUG dataset to empirically validate the effectiveness of our proposed HPLNet in maintaining performance across varying conditions. We selected all methods except FCN and U-net and conducted random image selection experiments on the validation set of the CHN6-CUG dataset, using weights trained on the DeepGlobe dataset.

The results of the generalisation experiments are presented in Figure 11. In the first image, both the comparison method and HPLNet failed to perform continuous segmentation of the trees obscuring the black box. VMUnet-2 and our method segmented the largest area of road within the box, but in comparison, VMUnet-2 still had incorrect segmentation in the lower left corner. In the second image, other methods produced poor results. Dlinknet, CoANet, SDseg, and DeNet failed to segment the road within the black box, while RcfsNet, VMUnet-2, and CSWin-UNet segmented only a small portion. Our method, however, effectively segmented the number and shape of the roads.

The aforementioned generalization experiments provide compelling evidence for the robustness and superior generalization performance of our proposed HPLNet. Notably, while several competing methods exhibit sensitivity to distribution shifts, HPLNet consistently manifests a greater capacity to sustain accurate segmentation accuracy, particularly under the challenging conditions illustrated in the second image. Although HPLNet considers lightweight design, this further proves the feasibility and reliability of HPLNet in practical applications, especially in real-world scenarios where data distribution varies.

6 Conclusion

This paper proposes a lightweight hierarchical perception network to address the challenge of balancing model lightness and extraction accuracy in road extraction under resource-constrained scenarios. Our core contributions are reflected in three aspects:

HPLNet designs a collaborative architecture of shallow perception and deep perception modules. Shallow perception efficiently captures raw pixel details through striped convolution, while deep perception uses a lightweight attention mechanism to

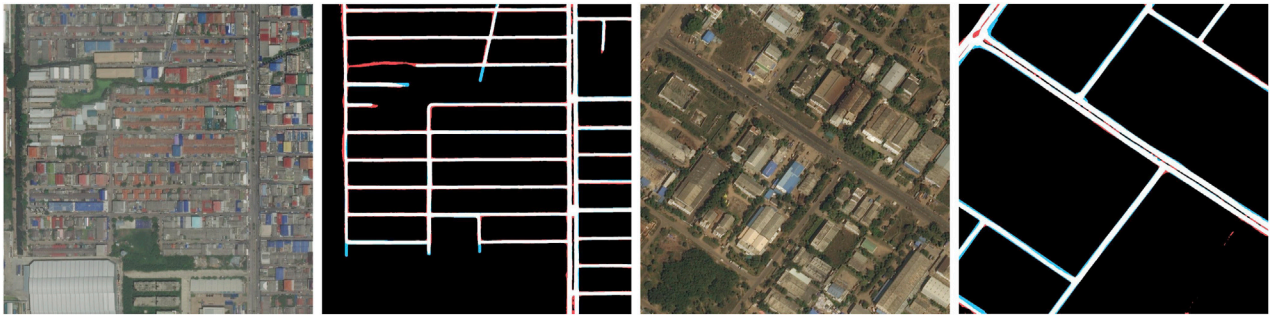

FIGURE 10
HPLNet visualisation results on DeepGlobe datasets.

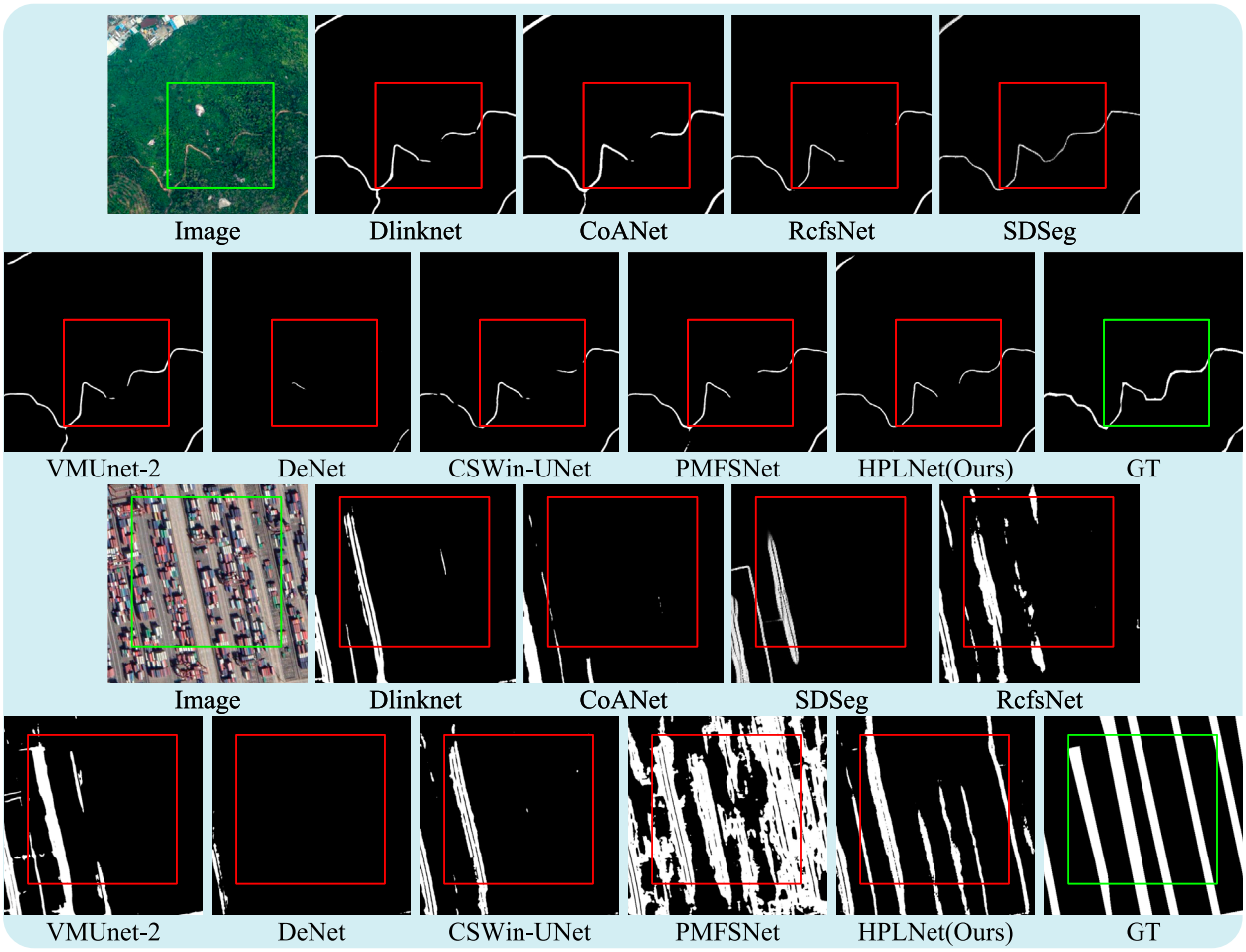

FIGURE 11
Qualitative comparison of generalisation experiments with other state-of-the-art methods in the CHN6-CUG dataset.

extract semantic information, achieving a good balance between lightweight network design and feature extraction accuracy under strict resource constraints. Second, to solve the problem of long-distance road extensions and local details being easily obscured, we cleverly combine long-distance information collection with local detail restoration. Specifically, shallow perception uses stripe convolution to capture long-distance road information from four directions to ensure accurate detail restoration, while deep perception introduces a lightweight spatial-channel attention mechanism to maintain network lightness while cooperatively retaining the global connectivity and local details of the road, balancing the extraction difficulties of the global and local.

Finally, to achieve deep coupling between shallow pixels and deep semantics, HPLNet adopts a hierarchical feature extraction strategy and introduces channel shift operations without additional parameter overhead, further ensuring the adequacy of shallow and deep feature extraction and achieving deep coupling of information.

Experimental results on mainstream benchmark datasets such as DeepGlobe and Massachusetts demonstrate that, compared to various advanced methods, the proposed HPLNet achieves a better balance between inference efficiency and extraction accuracy, and its effectiveness is further validated through generalization experiments on the CHN6 dataset.

Future work will focus on two directions: first, designing a more extreme lightweight network architecture that significantly reduces the number of model parameters and computational complexity while maintaining high accuracy; second, developing a road extraction method that can simultaneously take into account local details and global information to improve adaptability to complex and diverse remote sensing image scenes.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: http://deepglobe.org/challenge.html. https://www.cs.toronto.edu/~vmnih/data/.

Author contributions

SC: Validation, Conceptualization, Software, Investigation, Writing – original draft, Writing – review and editing, Visualization. QF: Formal Analysis, Funding acquisition, Project administration, Supervision, Writing – review and editing. LJ: Writing – review and editing, Investigation, Formal Analysis, Writing – original draft. XL: Software, Conceptualization, Methodology, Validation, Writing – review and editing. BG: Investigation, Writing – review and editing, Validation.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported in part by the National Key Research and Development Program of China under Grant 31400.

Conflict of interest

The 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.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Analysis Mach. Intell.* 39, 2481–2495. doi:10.1109/tpami.2016.2644615

Chaurasia, A., and Culurciello, E. (2017). "Linknet: exploiting encoder representations for efficient semantic segmentation," in 2017 IEEE visual communications and image processing (VCIP) *(IEEE)*. doi:10.1109/vcip.2017.8305148

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018). *Encoder-decoder with atrous separable convolution for semantic image segmentation*. Springer International Publishing, 833–851. doi:10.1007/978-3-030-01234-2_49

Chen, Z., Zhang, Y., Gu, J., Kong, L., Yang, X., and Yu, F. (2023). "Dual aggregation transformer for image super-resolution," in *Proceedings of the IEEE/CVF international conference on computer vision*, 12312–12321.

Cheng, K.-N., Ni, W., Zhang, H., Wu, J., Xiao, X., and Yang, Z. (2025). Ce-roadnet: a cascaded efficient road network for road extraction from high-resolution satellite images. *Remote Sens.* 17, 831. doi:10.3390/rs17050831

Cruz, G. G. L., Litonjua, A., Juan, A. N. P. S., Libatique, N. J., Tan, M. I. L., and Honrado, J. L. E. (2022). "Motorcycle and vehicle detection for applications in road safety and traffic monitoring systems," in *2022 IEEE global Humanitarian Technology Conference (GHTC)* (IEEE), 102–105. doi:10.1109/ghtc55712.2022.9910992

Cui, F., Feng, R., Wang, L., and Wei, L. (2021). "Joint superpixel segmentation and graph convolutional network road extration for high-resolution remote sensing imagery," in 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS *(IEEE)*, 2178–2181.

Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., et al. (2018). "Deepglobe 2018: a challenge to parse the earth through satellite images," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 172–181.

Dong, L. (2012). "The research on model framework of the trunk road network operation and environmental monitoring," in *2012 2nd international conference on remote sensing, environment and transportation engineering* (IEEE), 1–4. doi:10.1109/rsete.2012.6260785

Gui, L., Gu, X., Huang, F., Ren, S., Qin, H., and Fan, C. (2025). Road extraction from remote sensing images using a skip-connected parallel cnn-transformer encoder-decoder model. *Appl. Sci.* 15, 1427. doi:10.3390/app15031427

Guo, T., Gao, Y., Luo, F., Zhang, L., Du, B., and Gao, X. (2025). Denet: direction and edge co-awareness network for road extraction from high-resolution remote sensing imagery. *IEEE Trans. Intelligent Transp. Syst.* 26, 10236–10249. doi:10.1109/tits.2025.3546054

Han, X., Liu, Y., Liu, G., Lin, Y., and Liu, Q. (2023). Loanet: a lightweight network using object attention for extracting buildings and roads from uav aerial remote sensing images. *PeerJ Comput. Sci.* 9, e1467. doi:10.7717/peerj-cs.1467

Huang, Y., Wei, H., Yang, J., and Wu, M. (2021). "Damaged road extraction based on simulated post-disaster remote sensing images," in 2021 IEEE International Geoscience

and Remote Sensing Symposium IGARSS *(IEEE)*, 4684–4687. doi:10.1109/igarss47720.2021.9554812

Jamali, A., Roy, S. K., Li, J., and Ghamisi, P. (2024). Neighborhood attention makes the encoder of resunet stronger for accurate road extraction. *IEEE Geoscience Remote Sens. Lett.* 21, 1–5. doi:10.1109/lgrs.2024.3354560

Li, Y., Peng, B., He, L., Fan, K., Li, Z., and Tong, L. (2019). Road extraction from unmanned aerial vehicle remote sensing images based on improved neural networks. *Sensors* 19, 4115. doi:10.3390/s19194115

Li, C., Zeng, Q., Fang, J., Wu, N., and Wu, K. (2021). Road extraction in rural areas from high resolution remote sensing image using a improved Full Convolution Network. *Natl. Remote Sens. Bull.* 25, 1978–1988. doi:10.11834/jrs.20219209

Lian, R., Wang, W., Mustafa, N., and Huang, L. (2020a). Road extraction methods in high-resolution remote sensing images: a comprehensive review. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* 13, 5489–5507. doi:10.1109/jstars.2020.302359

Lian, R., Wang, W., Mustafa, N., and Huang, L. (2020b). Road extraction methods in high-resolution remote sensing images: a comprehensive review. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* 13, 5489–5507. doi:10.1109/jstars.2020.3023549

Lin, T., Chen, Z., Yan, Z., Yu, W., and Zheng, F. (2024). "Stable diffusion segmentation for biomedical images with single-step reverse process," in *International conference on medical image computing and computer-assisted intervention* (Springer), 656–666.

Liu, H., and Wang, Y. (2011). "The apply of urban design in the detailed planning of residential areas," in 2011 international conference on multimedia technology *(IEEE)*, 4164–4166. doi:10.1109/icmt.2011.6002786

Liu, B., Wu, H., Wang, Y., and Liu, W. (2015). Main road extraction from zy-3 grayscale imagery based on directional mathematical morphology and vgi prior knowledge in urban areas. *PLOS ONE* 10, e0138071. doi:10.1371/journal.pone.0138071

Liu, D., Zhang, J., Qi, Y., and Zhang, Y. (2023). A lightweight road detection algorithm based on multiscale convolutional attention network and coupled decoder head. *IEEE Geoscience Remote Sens. Lett.* 20, 1–5. doi:10.1109/lgrs.2023.3266054

Liu, X., Gao, P., Yu, T., Wang, F., and Yuan, R.-Y. (2025). Cswin-unet: transformer unet with cross-shaped windows for medical image segmentation. *Inf. Fusion* 113, 102634. doi:10.1016/j.inffus.2024.102634

Long, J., Shelhamer, E., and Darrell, T. (2015). "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Lu, X., Zhong, Y., and Zheng, Z. (2020). "A novel global-aware deep network for road detection of very high resolution remote sensing imagery," in IGARSS 2020-2020 IEEE international geoscience and remote sensing symposium *(IEEE)*, 2579–2582.

Mei, J., Li, R.-J., Gao, W., and Cheng, M.-M. (2021). Coanet: connectivity attention network for road extraction from satellite imagery. *IEEE Trans. Image Process.* 30, 8540–8552. doi:10.1109/tip.2021.3117076

Mnih, V. (2013). *Machine learning for aerial image labeling*. University of Toronto.

Mnih, V., and Hinton, G. E. (2010). *Learning to detect roads in high-resolution aerial images*. Berlin Heidelberg: Springer, 210–223. doi:10.1007/978-3-642-15567-3_16

Niu, X. (2006). A semi-automatic framework for highway extraction and vehicle detection based on a geometric deformable model. *ISPRS J. Photogrammetry Remote Sens.* 61, 170–186. doi:10.1016/j.isprsjprs.2006.08.004

Qi, H., Shi, J., Chen, J., Chi, C., and Shan, H. (2020). "Research on the complete design, construction and management of urban road in dalian city under the concept of "people-oriented traffic"," in *2020 5th international conference on electromechanical control technology and transportation (ICECTT)* (IEEE), 457–460. doi:10.1109/icectt50890.2020.00105

Qian, D., Wang, Y., Zhang, X., and Zhao, D. (2021). "Rationality evaluation of urban road network plan based on the ew-topsis method," in *2021 13th international conference on measuring technology and mechatronics automation (ICMTMA)* (IEEE), 840–844. doi:10.1109/icmtma52658.2021.00192

Ronneberger, O., Fischer, P., and Brox, T. (2015a). *U-Net: Convolutional networks for biomedical image segmentation*. Springer International Publishing, 234–241. doi:10.1007/978-3-319-24574-4_28

Ronneberger, O., Fischer, P., and Brox, T. (2015b). "U-net: convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18* (Springer), 234–241.

Saito, S., Yamashita, T., and Aoki, Y. (2016). Multiple object extraction from aerial imagery with convolutional neural networks. *Electron. Imaging* 28, 1–9. doi:10.2352/issn.2470-1173.2016.10.robvis-392

Seid, S., Zennaro, M., Libsie, M., Pietrosemoli, E., and Manzoni, P. (2020). "A low cost edge computing and lorawan real time video analytics for road traffic monitoring," in *2020 16th international conference on mobility, sensing and networking (MSN)* (IEEE), 762–767. doi:10.1109/msn50589.2020.00130

Shao, Z., Zheng, J., Yue, G., and Yang, Y. (2023). "Road traffic assignment algorithm based on computer vision," in *2023 international conference on integrated intelligence and communication systems (ICIICS)* (IEEE), 1–5. doi:10.1109/iciics59993.2023.10421615

Sultonov, F., Park, J.-H., Yun, S., Lim, D.-W., and Kang, J.-M. (2022). Mixer u-net: an improved automatic road extraction from uav imagery. *Appl. Sci.* 12, 1953. doi:10.3390/app12041953

Sun, T., Di, Z., Che, P., Liu, C., and Wang, Y. (2019). "Leveraging crowdsourced gps data for road extraction from aerial imagery," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7509–7518.

Sun, S., Yang, Z., and Ma, T. (2022). Lightweight remote sensing road detection network. *IEEE Geoscience Remote Sens. Lett.* 19, 1–5. doi:10.1109/lgrs.2022.3179400

Wan, Y., Hu, X., Zhong, Y., Ma, A., Wei, L., and Zhang, L. (2019). "Tailings reservoir disaster and environmental monitoring using the uav-ground hyperspectral joint observation and processing: a case of study in xinjiang, the belt and road," in Igarss 2019 - 2019 IEEE international geoscience and remote sensing symposium *(IEEE)*. doi:10.1109/igarss.2019.8898447

Wang, J., Qin, Q., Zhao, J., Ye, X., Qin, X., Yang, X., et al. (2015). "A knowledge-based method for road damage detection using high-resolution remote sensing image," in *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)* (IEEE), 3564–3567. doi:10.1109/igarss.2015.7326591

Wang, W., Yang, N., Zhang, Y., Wang, F., Cao, T., and Eklund, P. (2016). A review of road extraction from remote sensing images. *J. Traffic Transp. Eng. Engl. Ed.* 3, 271–282. doi:10.1016/j.jtte.2016.05.005

Wang, C., Xu, R., Xu, S., Meng, W., Wang, R., Zhang, J., et al. (2023a). Toward accurate and efficient road extraction by leveraging the characteristics of road shapes. *IEEE Trans. Geoscience Remote Sens.* 61, 1–16. doi:10.1109/tgrs.2023.3284478

Wang, R., Cai, M., Xia, Z., and Zhou, Z. (2023b). Remote sensing image road segmentation method integrating cnn-transformer and unet. *IEEE Access* 11, 144446–144455. doi:10.1109/access.2023.3344797

Wang, G., Yang, W., Ning, K., and Peng, J. (2024). Dfc-unet: a u-net-based method for road extraction from remote sensing images using densely connected features. *IEEE Geoscience Remote Sens. Lett.* 21, 1–5. doi:10.1109/lgrs.2023.3329803

Wei, Y., Zhang, K., and Ji, S. (2020). Simultaneous road surface and centerline extraction from large-scale remote sensing images using cnn-based segmentation and tracing. *IEEE Trans. Geoscience Remote Sens.* 58, 8919–8931. doi:10.1109/tgrs.2020.2991733

Wu, J., Wang, F., Li, X., Fan, J., Han, X., Zhou, Y., et al. (2018). "Disaster monitoring and emergency response services in China," in Igarss 2018 - 2018 IEEE international geoscience and remote sensing symposium *(IEEE)*, 3473–3476. doi:10.1109/igarss.2018.8519110

Wu, Q., Wang, S., Wang, B., and Wu, Y. (2022). Road extraction method of high-resolution remote sensing image on the basis of the spatial information perception semantic segmentation model. *Natl. Remote Sens. Bull.* 26, 1872–1885. doi:10.11834/jrs.20210021

Xiao, D., Yin, L., and Fu, Y. (2022). Open-pit mine road extraction from high-resolution remote sensing images using ratt-unet. *IEEE Geoscience Remote Sens. Lett.* 19, 1–5. doi:10.1109/lgrs.2021.3065148

Xu, Y., Liu, S., and Peng, Y. (2018). "Research and design of environmental monitoring and road lighting system based on the internet of things," in 2018 Chinese Automation Congress (CAC) *(IEEE)*, 1073–1078. doi:10.1109/cac.2018.8623501

Yager, N., and Sowmya, A. (2003). "Support vector machines for road extraction from remotely sensed images," in *International conference on computer analysis of images and patterns* (Springer), 285–292.

Yang, Z., Zhou, D., Yang, Y., Zhang, J., and Chen, Z. (2022). Transroadnet: a novel road extraction method for remote sensing images *via* combining high-level semantic feature and context. *IEEE Geoscience Remote Sens. Lett.* 19, 1–5. doi:10.1109/lgrs.2022.3171973

Yang, Z., Zhou, D., Yang, Y., Zhang, J., and Chen, Z. (2023a). Road extraction from satellite imagery by road context and full-stage feature. *IEEE Geoscience Remote Sens. Lett.* 20, 1–5. doi:10.1109/lgrs.2022.3228967

Yang, Z.-X., You, Z.-H., Chen, S.-B., Tang, J., and Luo, B. (2023b). Semisupervised edge-aware road extraction *via* cross teaching between cnn and transformer. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* 16, 8353–8362. doi:10.1109/jstars.2023.3310612

Yi, J., Shen, Z., Chen, F., Zhao, Y., Xiao, S., and Zhou, W. (2023). A lightweight multiscale feature fusion network for remote sensing object counting. *IEEE Trans. Geoscience Remote Sens.* 61, 1–13. doi:10.1109/tgrs.2023.3238185

Yuan, X., Shi, J., and Gu, L. (2021). A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Syst. Appl.* 169, 114417. doi:10.1016/j.eswa.2020.114417

Zhang, X., Li, T., and Zhao, X. (2023a). "Boosting single image super-resolution *via* partial channel shifting," in *Proceedings of the IEEE/CVF international conference on computer vision*, 13223–13232.

Zhang, X., Ma, X., Yang, Z., Liu, X., and Chen, Z. (2023b). A context-aware road extraction method for remote sensing imagery based on transformer network. *IEEE Geoscience Remote Sens. Lett.* 20, 1–5. doi:10.1109/lgrs.2023.3324644

Zhang, M., Yu, Y., Jin, S., Gu, L., Ling, T., and Tao, X. (2024). "Vm-unet-v2: rethinking vision mamba unet for medical image segmentation," in *International symposium on bioinformatics research and applications* (Springer), 335–346.

Zhao, S., Zhu, J., Lu, J., Ju, Z., and Wu, D. (2025). Lightweight human behavior recognition method for visual communication agv based on cnn-lstm. *Int. J. Crowd Sci.* 9, 133–138. doi:10.26599/ijcs.2024.9100014

Zhong, B., Dan, H., Liu, M., Luo, X., Ao, K., Yang, A., et al. (2025a). Ferdnet: High-resolution remote sensing road extraction network based on feature enhancement of road directionality. *Remote Sens.* 17, 376. doi:10.3390/rs17030376

Zhong, J., Tian, W., Xie, Y., Liu, Z., Ou, J., Tian, T., et al. (2025b). Pmfsnet: polarized multi-scale feature self-attention network for lightweight medical image segmentation. *Comput. Methods Programs Biomed.* 261, 108611. doi:10.1016/j.cmpb.2025.108611

Zhou, L., Zhang, C., and 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) *(IEEE)*, 192–1924. doi:10.1109/cvprw.2018.00034

Zhou, G., Chen, W., Gui, Q., Li, X., and Wang, L. (2021). Split depth-wise separable graph-convolution network for road extraction in complex environments from high-resolution remote-sensing images. *IEEE Trans. Geoscience Remote Sens.* 60, 1–15. doi:10.1109/tgrs.2021.3128033

Zhu, Q., Zhang, Y., Wang, L., Zhong, Y., Guan, Q., Lu, X., et al. (2021). A global context-aware and batch-independent network for road extraction from vhr satellite imagery. *ISPRS J. Photogrammetry Remote Sens.* 175, 353–365. doi:10.1016/j.isprsjprs.2021.03.016

Zhu, X., Huang, X., Cao, W., Yang, X., Zhou, Y., and Wang, S. (2024). Road extraction from remote sensing imagery with spatial attention based on swin transformer. *Remote Sens.* 16, 1183. doi:10.3390/rs16071183