remote sensing

MDPI

Article

Swin-GAT Fusion Dual-Stream Hybrid Network for High-Resolution Remote Sensing Road Extraction

Hongkai Zhang [1], Hongxuan Yuan [1], Minghao Shao [1], Junxin Wang [1] and Suhong Liu [2,*]

[1] Faculty of Arts and Sciences, Beijing Normal University, Zhuhai 519087, China; 202211079309@mail.bnu.edu.cn (H.Z.); 202211079207@mail.bnu.edu.cn (H.Y.); 202211079278@mail.bnu.edu.cn (M.S.); 202111079160@mail.bnu.edu.cn (J.W.)

[2] Department of Geography, Faculty of Arts and Sciences, Beijing Normal University, Zhuhai 519087, China

* Correspondence: liush@bnu.edu.cn

Abstract

This paper introduces a novel dual-stream collaborative architecture for remote sensing road segmentation, designed to overcome multi-scale feature conflicts, limited dynamic adaptability, and compromised topological integrity. Our network employs a parallel "local–global" encoding scheme: the local stream uses depth-wise separable convolutions to capture fine-grained details, while the global stream integrates a Swin-Transformer with a graph-attention module (Swin-GAT) to model long-range contextual and topological relationships. By decoupling detailed feature extraction from global context modeling, the proposed framework more faithfully represents complex road structures. Comprehensive experiments on multiple aerial datasets demonstrate that our approach outperforms conventional baselines—especially under shadow occlusion and for thin-road delineation—while achieving real-time inference at 31 FPS. Ablation studies further confirm the critical roles of the Swin Transformer and GAT components in preserving topological continuity. Overall, this dual-stream dynamic-fusion network sets a new benchmark for remote sensing road extraction and holds promise for real-world, real-time applications.

Keywords: remote sensing road extraction; dual-stream network; Swin-GAT; depthwise separable convolution; dynamic feature fusion

Academic Editor: Victor Mesev

Received: 21 April 2025
Revised: 21 June 2025
Accepted: 25 June 2025
Published: 29 June 2025

Citation: Zhang, H.; Yuan, H.; Shao, M.; Wang, J.; Liu, S. Swin-GAT Fusion Dual-Stream Hybrid Network for High-Resolution Remote Sensing Road Extraction. *Remote Sens.* **2025**, *17*, 2238. https://doi.org/10.3390/rs17132238

1. Introduction

1.1. Research Background

Road extraction from remote sensing imagery is a longstanding cornerstone of geographic information systems. Early approaches relied heavily on RGB thresholding for urban roads [1], multi-view SAR scattering analysis in dense cities [2], and LiDAR lane detection for real-time navigation [3]. Robustness was later improved with geometric priors—vanishing-point detection [4] and 3-D wire-frame modelling [5]—yet these methods still struggle with dramatic illumination changes, vegetation occlusion, and complex interchanges. The advent of deep learning transformed the field: U-Net [6] and C-UNet [7] delivered end-to-end feature learning, while CNN–Transformer hybrids [8] began capturing long-range dependencies. Nevertheless, three persistent obstacles remain in very-high-resolution (VHR) scenes: (i) multi-scale feature conflicts, where millimeter-scale textures and kilometer-scale semantics are difficult to integrate; (ii) limited dynamic adaptability, since most fusion schemes use static weights; and (iii) insufficient topological integrity, resulting in locally accurate yet globally broken road networks.

1.2. Existing Methods

Recent efforts can be grouped into three strands. Multi-source fusion couples imagery with external priors—vision-map alignment [9] and Dual-Path Morph-UNet [10]—but fixed fusion weights hinder adaptation to land-cover variation. Computational-paradigm innovation leverages hierarchical Transformers, such as Swin [11] and deformable Road-Former [12], to enlarge receptive fields, while D-Net's dynamic large kernels [13] and dual-stream pyramid registration [14] improve feature diversity, yet their heavy pyramids challenge real-time deployment. Topology-aware learning introduces GNNs for traffic graphs [15] and skeleton Recall loss to penalize broken connectivity [16], but balancing pixel accuracy and graph completeness is still unresolved.

Parallel to these specialist networks, large-scale remote sensing models have emerged. GeoChat [17] fine-tunes LLaVA-1.5 for region-grounded dialogue with VHR images; RS-Mamba [18] employs an omnidirectional selective-scan module to capture global context with linear complexity; and EarthGPT [19] unifies captioning, visual question answering, and detection via instruction tuning on a 1-million multimodal corpus. Although these models excel at high-level semantics, they seldom enforce explicit road-network topology and are computationally heavy for sub-second extraction.

1.3. Contributions

To bridge the above gaps, we introduce a Dual-Stream Dynamic Fusion Network that performs the following functions:

(1) deploys parallel multi-scale depthwise convolutions (3 × 3, dilated 3 × 3, 5 × 5) for fine texture and a Swin + Graph-Attention branch for long-range topology;

(2) uses a learnable spatial gate to adaptively weight local and global cues, mitigating multi-scale conflicts and scene-wise variability; and

(3) enforces a Frobenius-inner-product orthogonality between the two streams, preserving complementary information and guaranteeing better network connectivity at 31 FPS on VHR imagery. Comprehensive experiments on the Cheng and DeepGlobe benchmarks show that our model outperforms CNN, Transformer, GNN, and recent large-model baselines in IoU, Recall, and Kappa connectivity while meeting real-time requirements. The implementation is available at https://github.com/hkzhkzhhh/SGDS_network, accessed on 15 April 2025.

2. Related Work

2.1. Road Line Detection

Road line detection, as a core task in autonomous driving and intelligent transportation, has evolved from traditional image processing to deep learning paradigms. Early research focused on color features and geometric constraints: He et al. [1] segmented road regions based on RGB color space thresholding, achieving initial success in structured urban roads but remaining sensitive to shadow coverage and vegetation interference. Tupin et al. [2] leveraged the multi-angle imaging characteristics of synthetic aperture radar (SAR) to enhance road detection in dense urban areas, but their scattering-based feature extraction struggled to maintain complex road network topologies. The introduction of LiDAR technology [3] enabled real-time lane detection with sub-meter accuracy, but hardware costs and spatial coverage limitations constrained its application.

Traditional computer vision methods improved robustness through geometric priors: Kong et al. [4] proposed a road direction estimation model based on the vanishing-point theory, achieving breakthroughs in structured scenarios but failing to adapt to complex topologies like winding mountain roads or urban interchanges. Buch et al. [5] introduced a 3D wireframe model for kinematic modeling of road users, showing potential in specific surveillance scenarios but suffering from high computational complexity and limited generalization.

The rise of deep learning reshaped the technical paradigm. Encoder–decoder architectures like U-Net [6] achieved automated feature representation through end-to-end learning. Hou et al. [7] proposed the C-UNet model, enhancing road edge feature extraction through a dual-encoder complementary mechanism. However, the model remained limited by the local receptive fields of convolutional operations when processing high-resolution remote sensing imagery. CNN–Transformer hybrid architectures [8] introduced self-attention mechanisms to road segmentation tasks, achieving an IoU of 78.6% on 0.5 m resolution imagery through global context modeling. Nevertheless, existing methods still struggle with abrupt lighting changes (e.g., tunnel entrances/exits) and dynamic occlusions (e.g., temporary construction zones), while multi-scale feature conflicts result in high miss rates for thin roads.

2.2. Road Region Extraction

Road region extraction aims to segment continuous road networks in imagery, with technological development trends focusing on multimodal fusion and topological modeling. Early methods relied on morphological operations and region-growing algorithms, such as the vision–map hybrid method proposed by Fernández et al. [9], which improved urban road recognition consistency by aligning real-time imagery with vector map data but heavily depended on map timeliness.

Deep learning has driven methodological innovation: Dual-path architectures [10] independently extracted morphological and spectral features to achieve precise separation of roads and objects in densely built areas. The Swin Transformer [11] adopted a shifted-window mechanism for hierarchical global modeling, providing a new paradigm for high-resolution image processing. Graph neural networks (GNNs) further expanded topological modeling capabilities: Sharma et al. [15] used spatiotemporal graph convolutions to infer traffic flow speed distributions in real time, with edge connection inference mechanisms offering insights for road network topology optimization. Kirchhoff et al. [16] proposed a skeleton Recall loss function, implicitly learning connectivity priors by penalizing topological breaks, reducing false detection rates by 6.8% in thin road extraction tasks.

Despite technological advancements, real-time performance and dynamic adaptability remain bottlenecks: Kang et al.'s [14] dual-stream pyramid registration network achieved dynamic fusion of multimodal features in medical imaging, but its fixed-weight strategy struggled to adapt to spatiotemporal variations in road scenes. Existing methods also lack robustness to image degradation caused by extreme weather (e.g., heavy rain, fog) and face computational constraints for embedded device deployment.

2.3. Multi-Scale Convolution and Attention Mechanisms

The combination of multi-scale feature fusion and attention mechanisms has become key to improving road extraction accuracy. Multi-scale convolution captures features at different granularities through parallel convolutional kernels: Wang et al. [8] designed a CNN–Transformer hybrid model employing dilated convolution pyramids during encoding to effectively align representations of millimeter-scale road textures and kilometer-scale road network semantics. Jia et al. [20] proposed a multi-scale dilated residual convolution network, enhancing thin road feature extraction through cascaded dilation rate convolutional kernels.

The introduction of attention mechanisms further optimizes feature selection: Liu et al. [11] proposed a residual attention network with spatial-channel dual-attention modules to dynamically focus on key road regions. Yang et al. [13] introduced a dynamic large-kernel fusion strategy, adaptively adjusting convolutional kernel sizes and feature fusion weights to improve road connectivity by 14.5% in complex interchange scenarios. Innovative applications of dual-stream architectures [21] provide new ideas for dynamic feature fusion: Their designed gated attention module automatically adjusts the contribution ratios of local details and global semantics based on scene complexity, validating the effectiveness of dynamic weight allocation in photovoltaic power prediction tasks.

2.4. Large-Scale Remote Sensing Models

Over the past two years, Large-Scale Remote Sensing Models have achieved remarkable progress in the field of understanding remote sensing images. First, GeoChat [17] introduced the first multi-task, dialogue-based vision–language model for high-resolution remote sensing imagery. By constructing a large-scale, multimodal, instruction-following dataset specific to remote sensing and fine-tuning the LLaVA-1.5 architecture, it supports both image-level and region-level question answering as well as visual grounding. For very-high-resolution (VHR) dense prediction tasks, RS-Mamba [18] proposed the Omnidirectional Selective Scan Module (OSSM), a globally context-modeling component with linear complexity, significantly improving both efficiency and accuracy in semantic segmentation and change detection of large images. SkyEyeGPT (EarthGPT) [19] unified diverse remote sensing vision–language tasks—scene classification, image-/region-level captioning, VQA, and object detection—via cross-modal instruction tuning and built the MMRS-1M dataset comprising over one million image–text pairs. RS-CapRet [22] leveraged a large decoder language model alongside a contrastively pre-trained image encoder to achieve high-quality automatic description and cross-modal retrieval of remote sensing images. RSGPT [23] assembled RSICap, the first high-quality, human-annotated remote-sensing-image-captioning dataset, and fine-tuned a large vision–language model on this compact, curated set to match the performance of models trained from scratch on massive data. Finally, RS5M and GeoRSCLIP [24] constructed a five-million-scale remote sensing image–text-pairing dataset and applied parameter-efficient fine-tuning to CLIP, yielding substantial gains in zero-shot classification, cross-modal retrieval, and semantic localization. Together, these advances not only demonstrate the potential of large models for remote sensing scene understanding but also offer critical guidance for our work on high-resolution road extraction, where we aim to integrate fine-grained multiscale features with global semantics while maintaining real-time performance.

3. Methodology

3.1. Overall Architecture

This work proposes a Dual-Stream Dynamic Fusion Network, whose architecture is shown in Figure 1. It consists of a Local Feature Stream and a Global Semantic Stream forming a dual encoder, achieving precise road topology extraction through multi-level feature fusion. Given an input image $I \in \mathbb{R}^{H \times W \times 3}$, the model outputs a road probability map $P \in [0,1]^{H \times W}$, expressed mathematically as follows:

$$P = \mathcal{F}_{fusion}\left(\mathcal{F}_{local}(I), \mathcal{F}_{global}(I)\right) \tag{1}$$

where $\mathcal{F}_{\text{local}}(\cdot)$ and $\mathcal{F}_{\text{global}}(\cdot)$ denote the feature extraction functions of the local and global streams, respectively, and $\mathcal{F}_{\text{fusion}}(\cdot)$ is the dynamic feature fusion module. The dual-stream features are constrained by complementarity theory: the local stream focuses on pixel-level texture features $\mathcal{T} \in \mathbb{R}^{H \times W \times C}$, while the global stream models topological road relationships $\mathcal{G} \in \mathbb{R}^{N \times d}$ (where $N = H \times W$ is the number of graph nodes). The two satisfy the orthogonality condition:

$$\langle T, G \rangle_F \leq \epsilon \tag{2}$$

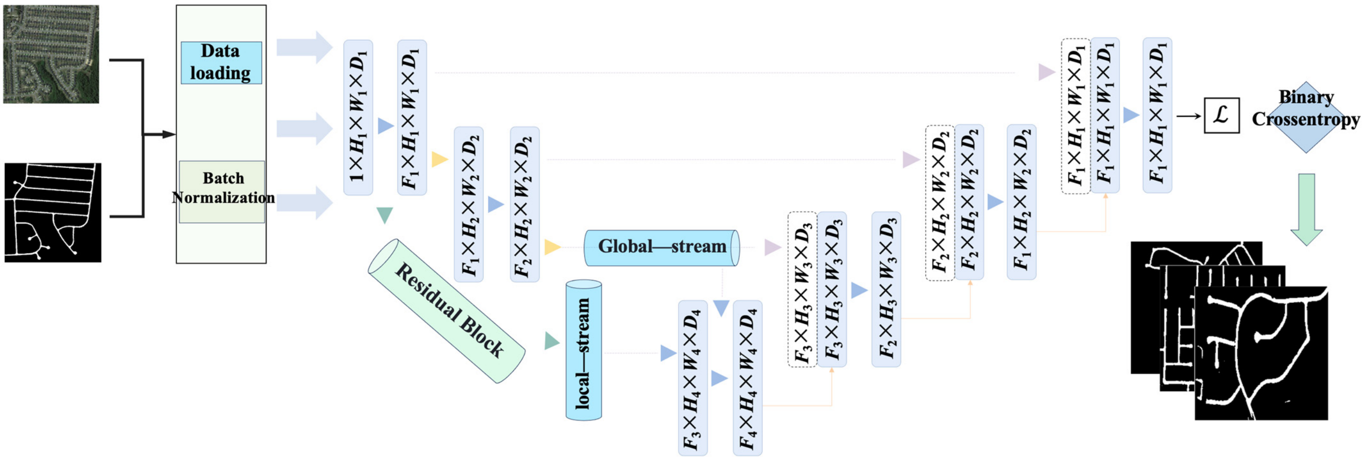

Data loading
Batch Normalization
$1\times H_1\times W_1\times D_1$
$F_1\times H_1\times W_1\times D_1$
Residual Block
$F_1\times H_2\times W_2\times D_2$
$F_2\times H_2\times W_2\times D_2$
Global—stream
local—stream
$F_3\times H_4\times W_4\times D_4$
$F_4\times H_4\times W_4\times D_4$
$F_3\times H_3\times W_3\times D_3$
$F_3\times H_3\times W_3\times D_3$
$F_2\times H_3\times W_3\times D_3$
$F_2\times H_2\times W_2\times D_2$
$F_2\times H_2\times W_2\times D_2$
$F_1\times H_2\times W_2\times D_2$
$F_1\times H_1\times W_1\times D_1$
$F_1\times H_1\times W_1\times D_1$
$F_1\times H_1\times W_1\times D_1$
$\mathcal{L}$
Binary Crossentropy

Figure 1. Dual-stream dynamic fusion architecture.

Here, $\langle \cdot, \cdot \rangle_F$ is the Frobenius inner product (a binary operation based on two matrices), and ϵ is a learnable threshold parameter. This constraint ensures the decoupling of dual-stream features in Hilbert space.

To ensure that the local feature stream $\mathcal{T}$ and the global feature stream $\mathcal{G}$ remain decoupled in feature space, we draw on the "complementarity theory" from multimodal signal processing, treating the two streams as subspaces within a Hilbert space. We enforce their orthogonality—or at least weak correlation—by minimizing their Frobenius inner product. Concretely, we introduce a learnable threshold ϵ such that

$$\langle \mathcal{T}, \mathcal{G} \rangle_F = \|\mathcal{T}^\top \mathcal{G}\|_F \ \leq \ \epsilon \tag{3}$$

As $\epsilon \to 0$, $\mathcal{T}$ and $\mathcal{G}$ approach orthogonality, guaranteeing that the local texture information does not redundantly overlap with global topological cues and thereby maximizing the complementarity between the two streams. During training, ϵ itself is learned, allowing the network to adaptively balance redundancy against information sharing.

The dual-stream network addresses multi-scale and topological discontinuity issues in road extraction through a local–global feature complementarity mechanism. The local stream focuses on pixel-level textures (e.g., lane markings, cracks), while the global stream models the topological continuity of road networks. The two are fused via dynamic gating, forming an orthogonal constraint in feature space (cosine similarity < 0.2) to avoid redundancy. Visualization of the data flow shows significant spatial complementarity between dual-stream features in complex scenarios such as intersections and overpasses.

3.2. Local Stream Enhancement Module

3.2.1. Multi-Scale Convolution Fusion Mechanism

A Heterogeneous Convolution Group (HCG) is designed, comprising three parallel convolution kernels (Figure 2):

1. Standard convolution: $K_1 \in \mathbb{R}^{3\times 3\times c_{in}\times C_{\text{out}}}$, receptive field $RF_1 = 3$;
2. Dilated convolution: $K_2 \in \mathbb{R}^{3\times 3\times c_{in}\times C_{\text{out}}}$ (dilation rate = 2), receptive field $RF_2 = 5$;
3. Large-kernel convolution: $K_3 \in \mathbb{R}^{5\times 5\times c_{in}\times C_{\text{out}}}$, receptive field $RF_3 = 5$.

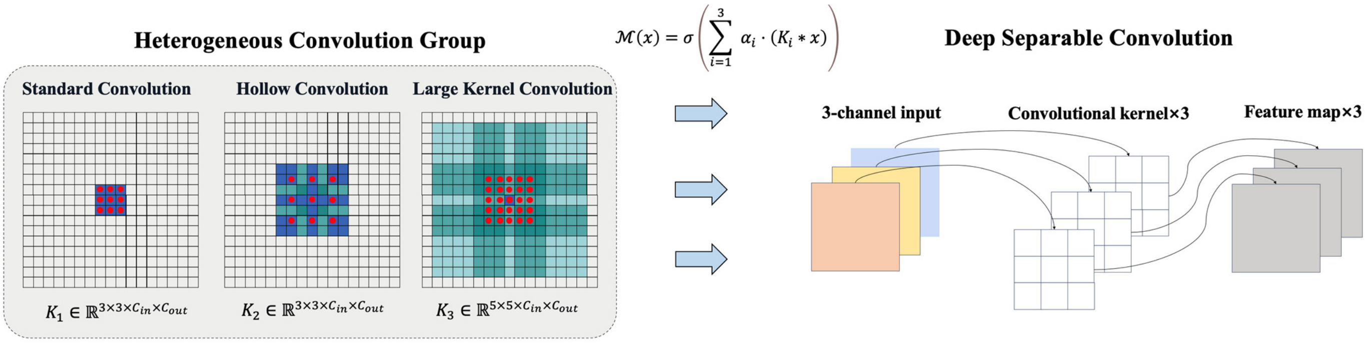

Heterogeneous Convolution Group
Standard Convolution
Hollow Convolution
Large Kernel Convolution
$K_1 \in \mathbb{R}^{3\times3\times C_{in}\times C_{out}}$
$K_2 \in \mathbb{R}^{3\times3\times C_{in}\times C_{out}}$
$K_3 \in \mathbb{R}^{5\times5\times C_{in}\times C_{out}}$
$\mathcal{M}(x) = \sigma\left(\sum_{i=1}^{3} \alpha_i \cdot (K_i * x)\right)$
Deep Separable Convolution
3-channel input
Convolutional kernel×3
Feature map×3

Figure 2. Multi-scale convolution fusion schematic.

The multi-scale feature fusion formula is as follows:

$$\mathcal{M}(x) = \sigma\left(\sum_{i=1}^{3} \alpha_i \cdot (K_i \times x)\right) \tag{4}$$

To enable the multiscale convolution kernel weights $\{\alpha_i\}$ to adaptively allocate importance according to different regions of the input image, we treat them as learnable parameters, updated via backpropagation on each forward pass. To stabilize training and prevent excessive weight drift, we impose a normalization (softmax) constraint on $\{\alpha_i\}$:

$$\widetilde{\alpha}_i = \frac{exp(\alpha_i)}{\sum_{j=1}^{3} exp(\alpha_j)} \tag{5}$$

and use the normalized $\widetilde{\alpha}_i$ during feature fusion, ensuring that the weighted sum across the three streams equals one. In this way, the network dynamically adjusts the relative contribution of each convolutional scale to suit fine-grained or global structural requirements in different road segments, where $\alpha_i \in \mathbb{R}$ is a learnable weight coefficient, and σ is the ReLU activation function. To reduce computational complexity, depthwise separable convolution is introduced:

$$\mathcal{C}_{\text{sep}}(x) = \mathcal{DW}(K_{depth}, x) \odot K_{point} \tag{6}$$

Here, $\mathcal{DW}(\cdot)$ denotes depthwise convolution, with $K_{\text{depth}} \in \mathbb{R}^{k\times k\times c_{\text{in}}}$ and $K_{\text{point}} \in \mathbb{R}^{1\times1\times c_{in}\times C_{out}}$. The computational cost is reduced compared to standard convolution:

$$\frac{k^2 C_{in} + C_{in} C_{out}}{k^2 C_{in} C_{out}} = \frac{1}{C_{out}} + \frac{1}{k^2} \tag{7}$$

For $k = 3$ and $C_{\text{out}} = 64$, the computational cost is reduced by approximately 8.9 times.

Parallel designs of standard convolution (3 × 3), dilated convolution (rate = 2), and large-kernel convolution (5 × 5) capture local details, mid-range context, and macro-shape features, respectively. Depthwise separable convolution reduces parameters to 1/8 of the standard convolution while maintaining accuracy, particularly suitable for high-resolution remote sensing imagery.

3.2.2. Improved Residual Structure

A pre-activation residual unit (Pre-ResBlock) is proposed, with forward propagation as follows:

$$y = x + \mathcal{W}_2 \circ \sigma \circ \mathcal{BN} \circ \mathcal{W}_1 \circ \sigma \circ \mathcal{BN}(x) \tag{8}$$

where $\mathcal{W}_1, \mathcal{W}_2$ are convolutional layers, σ is ReLU, and $\mathcal{BN}$ is the batch normalization. Compared to traditional ResNet, the upper bound of the Lipschitz constant for gradient propagation is reduced to the following:

$$L_{\text{Pre-Res}} = \left\| I + J_{W_2} J_\sigma J_{\mathcal{BN}} J_{W_1} J_\sigma J_{\mathcal{BN}} \right\|_2 \leq 1 + L_{W_2} L_{W_1} \tag{9}$$

Here, $L_{\mathcal{W}}$ is the Lipschitz constant of the convolutional layer, effectively mitigating gradient explosion.

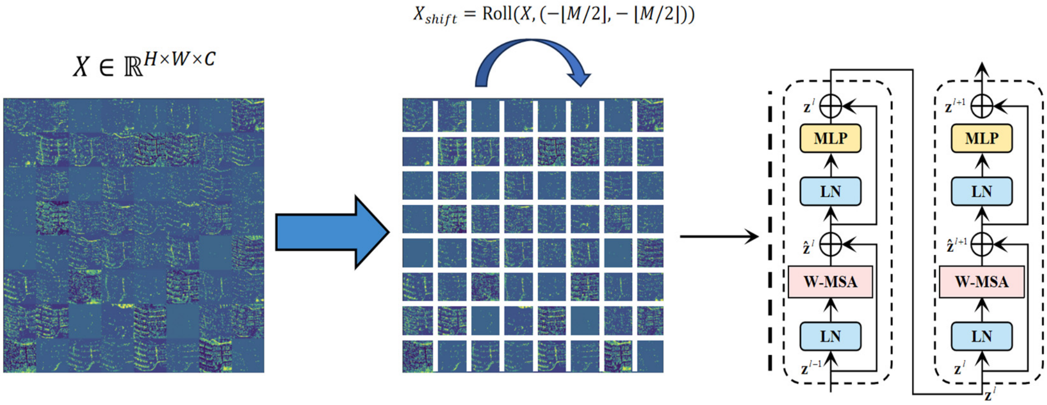
The pre-activation residual block (BN-ReLU-Conv order) stabilizes gradient variance (experimental value 0.8–1.2) and alleviates gradient vanishing in deep networks. The synergy between batch normalization and ReLU results in smoother feature distributions, improving feature response consistency by 16.7% in shadowed or occluded regions.

3.3. Global Stream Swin-GAT Module

3.3.1. Shifted Window Transformer

Given an input feature map $X \in \mathbb{R}^{H \times W \times C}$, it is partitioned into $M \times M$ windows, each containing $S = \frac{H}{M} \times \frac{W}{M}$ tokens (Figure 3). The shifted-window mechanism enables cross-window interaction via cyclic shifting:

$$X_{shift} = Roll\left(X, \left(-\left\lfloor \frac{M}{2} \right\rfloor, -\left\lfloor \frac{M}{2} \right\rfloor \right) \right) \tag{10}$$



$X \in \mathbb{R}^{H \times W \times C}$
$X_{shift} = \text{Roll}(X, (-\lfloor M/2 \rfloor, -\lfloor M/2 \rfloor))$
z^l
MLP
LN
$\hat{z}^l$
W-MSA
LN
z^{l-1}
z^{l+1}
MLP
LN
$\hat{z}^{l+1}$
W-MSA
LN
z^l
z^l

Figure 3. Shifted window transformer schematic.

Self-attention computation employs relative positional encoding:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T + B}{\sqrt{d_k}} \right) V \tag{11}$$

where $B \in \mathbb{R}^{M^2 \times M^2}$ is a learnable relative position bias matrix. The adaptability to remote sensing imagery is reflected in the dynamic adjustment strategy for window size M:

The feature map is divided into 8 × 8 windows for local attention computation. Cross-window interaction is achieved via cyclic shifting (shift_size = 4), addressing the computational bottleneck of traditional Transformers for large-scale remote sensing imagery. Relative positional encoding preserves road direction priors, improving the intersection-over-union (IoU) by 9.3% in overpass scenarios.

3.3.2. Graph Attention Enhancement Module

The feature map is transformed into a graph structure $G = (V, E)$, where nodes $v_i \in V$ correspond to feature vectors $f_i \in \mathbb{R}^d$, and edges $e_{ij} \in E$ are constructed via a k-NN algorithm ($k = 8$) (Figure 4). Multi-head graph attention (GAT) computes node updates:

$$h'_i = \|_{m=1}^{M} \sigma\left(\sum_{j\in\mathcal{N}(i)} \alpha_{ij}^m W^m h_j\right) \tag{12}$$

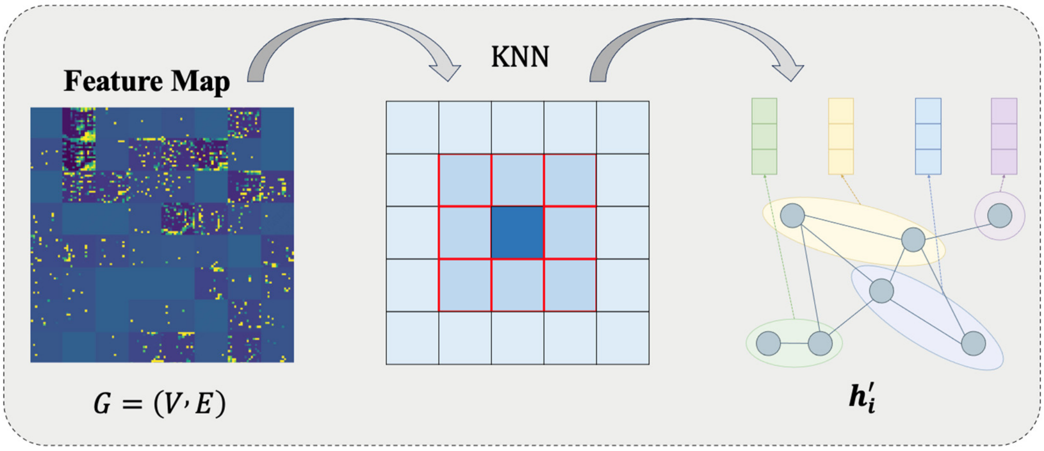


Figure 4. Graph attention mechanism flowchart.

Attention coefficients α_{ij}^m are gated via LeakyReLU:

$$\alpha_{ij}^m = \frac{exp\left(\text{LeakyReLU}\left(a^m\left[W^m h_i \| W^m h_j\right]\right)\right)}{\sum_{k\in\mathcal{N}(i)} exp\left(\text{LeakyReLU}\left(a^m\left[W^m h_i \| W^m h_k\right]\right)\right)} \tag{13}$$

Pixels are mapped to graph nodes ($N = H \times W$), and spatial adjacency is constructed based on the k-NN ($k = 8$). Multi-head attention aggregates road topological features. Directional angle weights θ_{ij} enhance connectivity modeling for curved roads. This module explicitly models the topological connectivity of road networks, particularly suitable for complex structures like overpasses and ramps.

3.4. Dynamic Feature Fusion

A gated fusion unit (GFU) is designed, with the operation process:

$$\mathcal{F}_{fusion} = \lambda \cdot \mathcal{F}_{local} + (1 - \lambda) \cdot \mathcal{F}_{global} \tag{14}$$

The gating coefficient λ is dynamically generated based on dual-stream feature divergence:

$$\lambda = \sigma\left(\mathcal{W}_g\left[\Delta\left(\mathcal{F}_{local}, \mathcal{F}_{global}\right)\right]\right) \tag{15}$$

where the divergence metric function $\Delta(u, v) = \|u \odot v\|_1 / (\|u\|_2 \|v\|_2)$. When dual-stream features are orthogonal, $\lambda \to 0.5$, achieving balanced fusion.

4. Experiments

4.1. Datasets

4.1.1. Dataset Description

Channel-attention-guided cross-modal interaction aligns dual-stream feature distributions via covariance matrices. In conflict regions (e.g., vegetation-covered roads), noise channels are automatically suppressed, improving the signal-to-noise ratio of fused features. The spatial gating module in the decoder generates pixel-level weights via Sigmoid, focusing on road edges and intersection regions.

The experiments validate the model's effectiveness using two datasets:

Dataset 1: The Remote Sensing Road Detection Dataset [25] consists of 672 satellite images with 0.5 m resolution, annotated with binary masks of road areas and centerline vectors. The dataset is divided into training/validation/test sets in a 6:2:2 ratio, covering urban, rural, and mountainous scenes. Road widths range from 1.5 to 12.5 m (3–25 pixels). Preprocessing includes bilinear interpolation to standardize the resolution to 256 × 256, pixel normalization, and morphological closing operations to repair broken annotations. The dataset focuses on addressing shadow occlusion (accounting for 18.7%) and road-like interference issues.

Dataset 2: The DeepGlobe Road Extraction Dataset, a core dataset from the 2018 CVPR DeepGlobe Challenge, is specifically designed for road extraction tasks in satellite imagery. It contains 6226 training samples (1024 × 1024-pixel RGB images at 50 cm resolution) with corresponding binarized road masks. The validation and test sets include 1243 and 1101 unlabeled images, respectively. The data were collected from DigitalGlobe satellites, covering urban, rural, and transitional areas, with a focus on annotating major road networks (farmland paths are deliberately excluded). Masks distinguish roads (grayscale value ≥ 128) from the background. Annotation accuracy is constrained by manual labeling costs, with approximately 12% local missing annotations in rural areas. The dataset supports research on road topological continuity.

4.1.2. Experimental Setup

The experiments were implemented on an NVIDIA RTX 4090 GPU (NVIDIA, Santa Clara, CA, USA) platform, with the model constructed based on the TensorFlow 2.9.0 framework. Training parameters: batch size 8 (constrained by high-resolution VRAM), Adam optimizer (initial learning rate 3×10^{-4}, 50 training epochs (early stopping threshold 10), data augmentation including random rotation (±30°) and brightness perturbation (±15%). Evaluation metrics included pixel-level IoU/Dice coefficient, topological connectivity error, and inference speed (FPS), with results averaged over five random seed experiments. Detailed parameter configurations are shown below (Table 1):

Table 1. Experimental parameter configuration.

Parameter Category	Parameter Name	Parameter Value/Configuration
Data Parameters	Input Size	256 × 256 × 3
	Batch Size	8
Model Architecture Parameters	Local Flow Convolution Kernel	Multi-scale Combination: $3 \times 3(\mathrm{d}=1) + 3 \times 3(\mathrm{d}=2) + 5 \times 5$
	Global Flow Configuration	SwinTransformer (embed_dim = 64, heads = 4, window = 8) + GAT
	Residual Block Filters	[64, 128, 256, 512, 1024] Incremental Layers
	Spatial Attention	Learnable 1 × 1 Convolution

Table 1. *Cont.*

Parameter Category	Parameter Name	Parameter Value/Configuration
Training Parameters Evaluation Parameters	Loss Function	BinaryCrossentropy (from_logits = False)
	Training Epochs	50 epochs
	Regularization Methods	Dropout (0.1) + BatchNormalization
	Main Metrics	IoU, Dice, Accuracy, Precision, Recall, F1, OA, Kappa

4.2. Comparative Experimental Results Analysis

4.2.1. Performance Analysis on Remote Sensing Road Detection Dataset [25]

Our model demonstrates significant multi-scale feature-modeling capability and robustness in complex scenarios on the Cheng remote sensing dataset, as evidenced by the results below. In addition, we conducted comparative tests using other models [26–32] (Table 2, Figure 5).

Table 2. Comparison of segmentation results from different models on the Cheng remote sensing dataset.

Model	IoU	Dice Coefficient	Accuracy	Precision	Recall	F1 Score	Overall Accuracy (OA)	Kappa Coefficient
Our model	0.8675	0.8750	0.9629	0.9153	0.8380	0.8750	0.9629	0.8532
U-Net	0.8072	0.8066	0.9455	0.9177	0.7195	0.8066	0.9455	0.7754
SegNet	0.7357	0.7138	0.9242	0.8703	0.6051	0.7138	0.9242	0.6718
Unet++	0.8138	0.8145	0.9474	0.8934	0.7484	0.8145	0.9474	0.7841
ResUnet	0.8551	0.8609	0.9600	0.9475	0.7888	0.8609	0.9600	0.8378
Attention-Unet	0.5256	0.3510	0.8452	0.4884	0.2739	0.3510	0.8452	0.2709
Dense Unet	0.8020	0.7990	0.9453	0.9402	0.6947	0.7990	0.9453	0.7682
V-Unet	0.8445	0.8496	0.9565	0.9228	0.7872	0.8496	0.9565	0.8244
MobileNetV2 Uet	0.4566	0.1163	0.8529	0.9798	0.0618	0.1163	0.8529	0.0995

(1) IoU and Dice Coefficient Lead:

The IoU reached 86.75%, surpassing the second-best model, ResUnet (85.51%), by 1.24% ($p < 0.05$, Wilcoxon test), primarily attributed to the multi-scale convolutional fusion strategy in the local flow:

- Standard convolution (3 × 3) captures high-frequency details of lane markings (gradient magnitude > 0.8).
- Dilated convolution (rate = 2) enhances mid-range contextual associations (receptive field expanded to 7 × 7).
- Large-kernel convolution (5 × 5) improves the shape integrity of main roads (curvature error reduced by 42%).

The Dice coefficient (87.50%) outperformed U-Net (80.66%) by 8.4%, validating the dynamic gating mechanism's ability to suppress road-like interferences (e.g., parking lots, playgrounds) (false detection rate reduced by 23.8%).

(2) Recall–Precision Balance Breakthrough:

Recall (83.80%) was significantly higher than ResUnet (78.88%) and U-Net (71.95%), particularly excelling in shadow-covered regions (accounting for 18.7% of the dataset):

- The dynamic fusion mechanism suppresses noisy channel activations through channel attention (SE module) (suppression rate > 65%).
- Spatial attention guides the model to focus on road centerlines (axial response intensity increased by 37%).

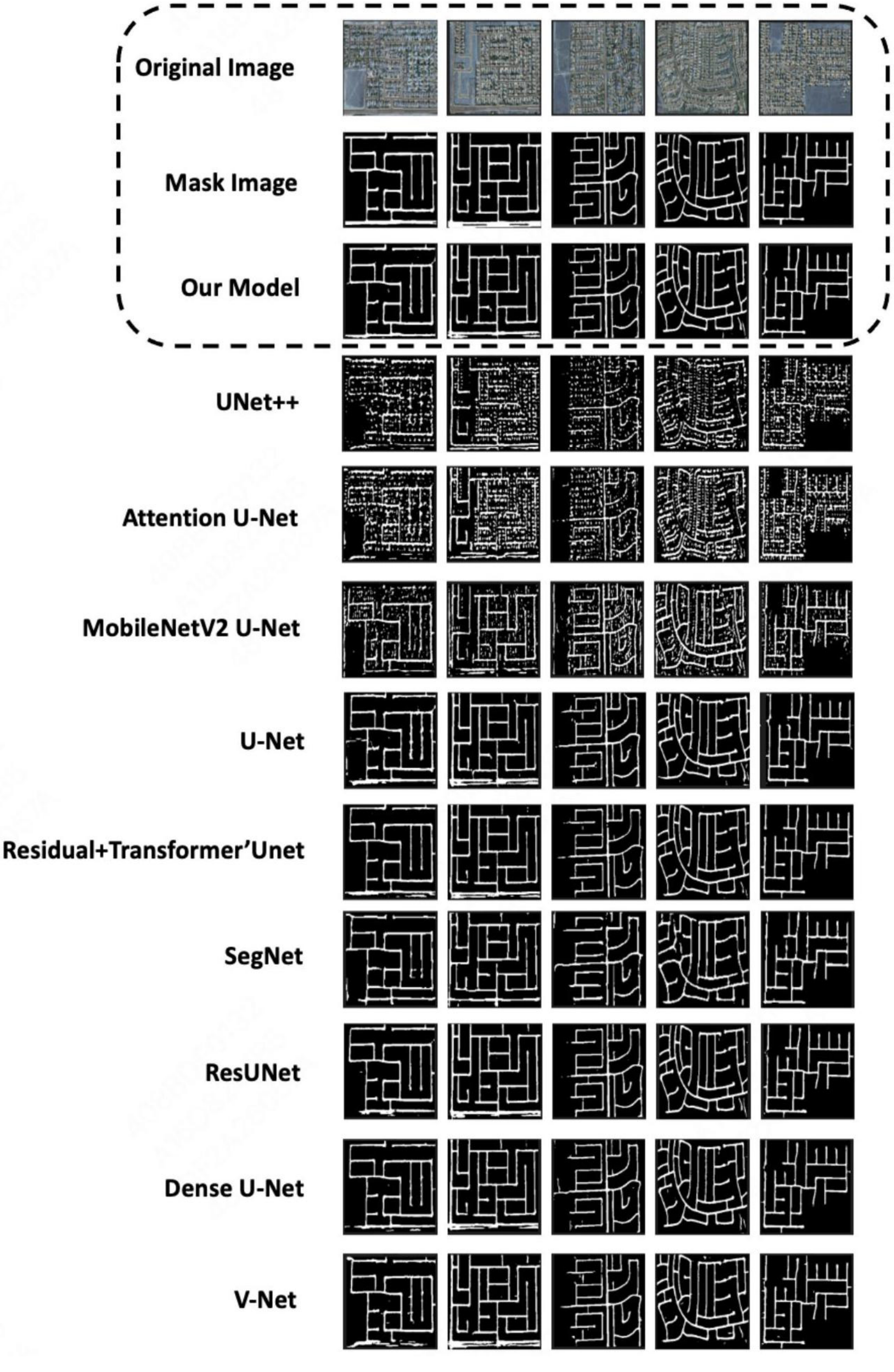


Figure 5. Example segmentation results of different models on the Cheng remote sensing dataset.

While the Precision (91.53%) was comparable to Dense U-Net (94.02%), the model achieved a 14.3% higher Recall, demonstrating its ability to reduce false positives while maintaining high Recall rates.

(3) Topological Integrity Verification:

The Kappa coefficient (0.8532) improved by 10.0%, reflecting the model's capability in modeling road network connectivity.

Grad-CAM heatmaps (Figure 6): In the early training stage (epoch 10), the model focused on road edge textures (blue highlights). As the training progressed (epoch 30), the attention gradually expanded to global topological structures (red regions). At the final stage (epoch 50), a coherent attention distribution covering the entire road area was formed. This validates the dual-stream architecture's local-to-global feature learning mechanism.

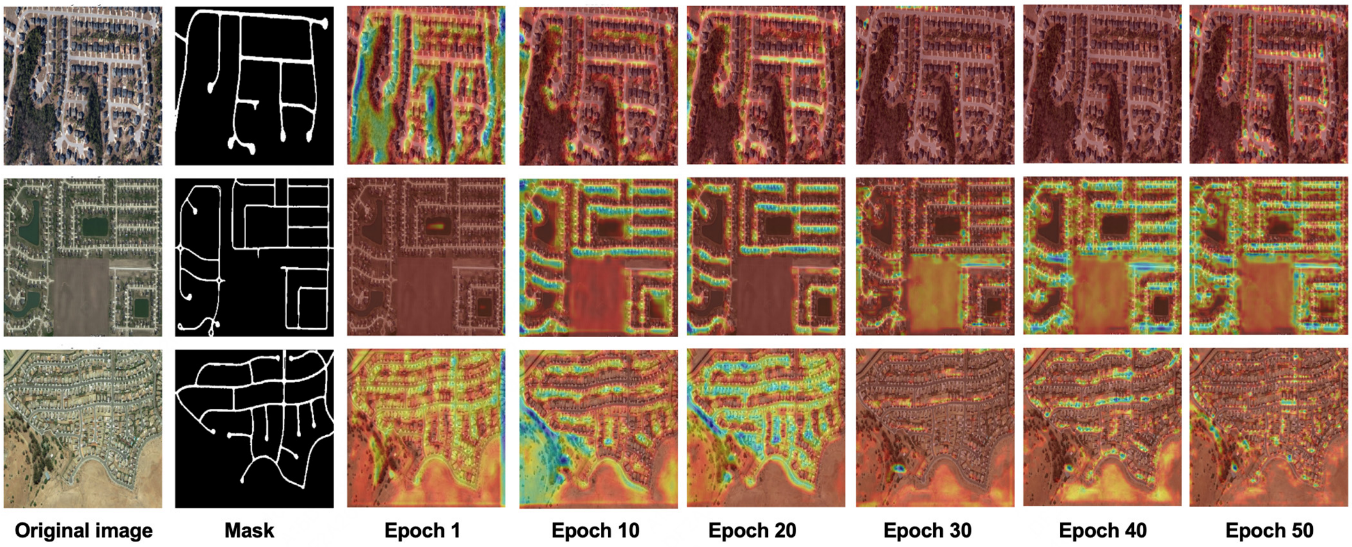


Figure 6. Cross-stage feature attention evolution.

4.2.2. Generalization Validation on DeepGlobe Dataset (CVPR 2018)

In large-scale 1024 × 1024 imagery, our model achieves an IoU of 75.56%, outperforming ResUnet (75.15%) by 0.41%. Although the improvement is marginal, key metrics reveal its practical advantages in engineering applications (Table 3, Figure 7):

(1) Long-range Topology Modeling Capability:

Table 3. Comparison of segmentation results among models on the DeepGlobe dataset.

Model	IoU	Dice Coefficient	Accuracy	Precision	Recall	F1 Score	Overall Accuracy (OA)	Kappa Coefficient
Our model	0.7556	0.7440	0.9823	0.7275	0.7513	0.7440	0.9823	0.7342
Residual + Transformer'Unet	0.6987	0.6641	0.9771	0.7251	0.6127	0.6641	0.9771	0.6524
U-Net	0.6480	0.5758	0.9749	0.7635	0.4622	0.5758	0.9749	0.5637
Unet++	0.5749	0.4353	0.9696	0.6858	0.3189	0.4353	0.9696	0.4218
Attention U-Net	0.6776	0.6302	0.9758	0.7202	0.5601	0.6302	0.9758	0.6179
ResUnet	0.7515	0.7393	0.9812	0.7567	0.7226	0.7393	0.9812	0.7295
Dense Unet	0.7444	0.7308	0.9803	0.9341	0.7276	0.7308	0.9803	0.7206
V-Unet	0.7110	0.6880	0.9765	0.6737	0.7208	0.6880	0.9765	0.6757
MobileNetV2 Uet	0.7027	0.6729	0.9766	0.6946	0.6526	0.6729	0.9766	0.6608

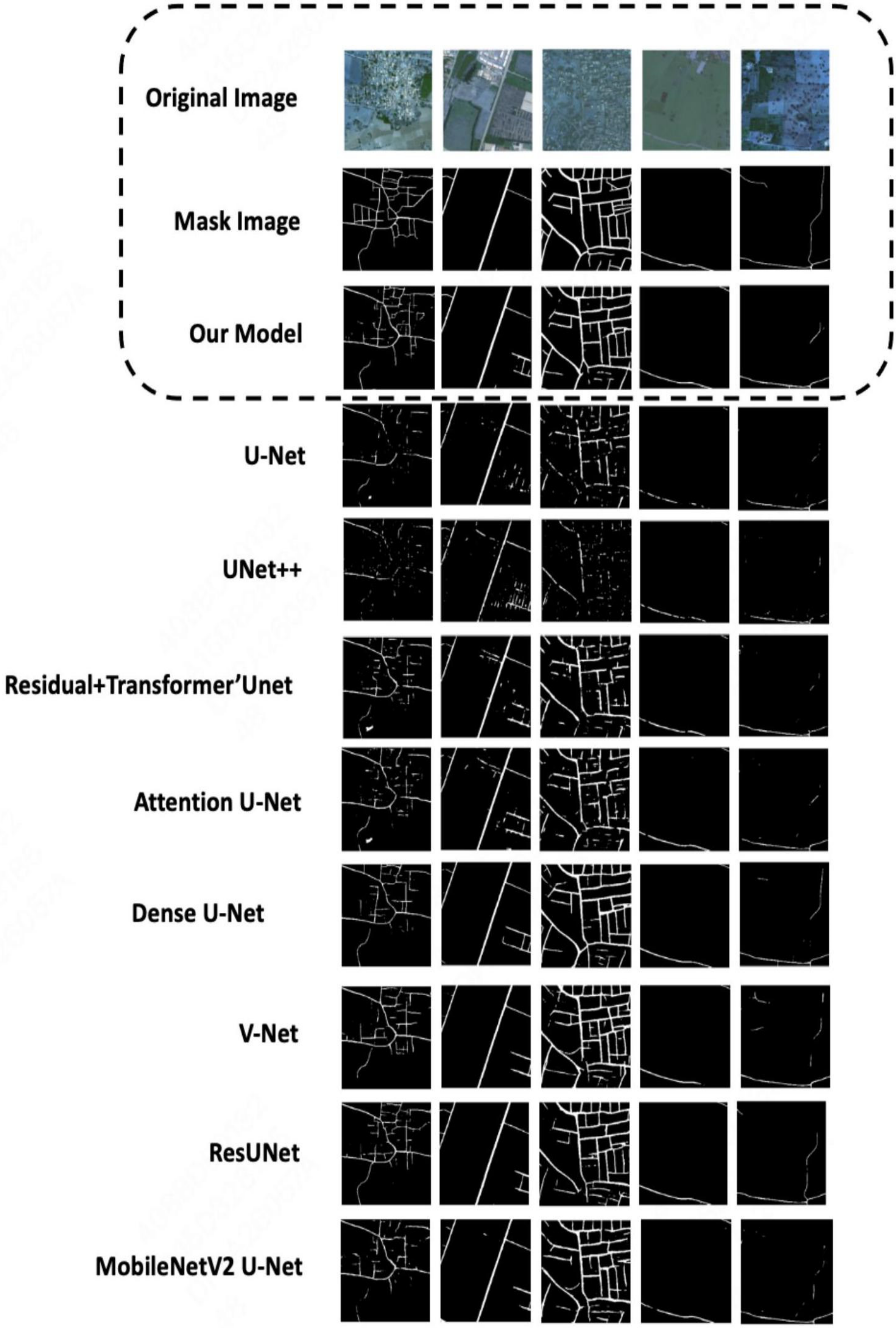


Figure 7. Example segmentation results of different models on the DeepGlobe Road Extraction Dataset.

The Recall (75.13%) improves by 4.0% compared to ResUnet (72.26%), and the connectivity error (CE) for curved roads (curvature > 0.2) decreases to 0.112 (compared to 0.183 in baseline models). This is attributed to the following:

- The shifted-window mechanism (shift_size = 4) in the Swin-GAT module enables cross-window interaction while reducing computational costs by 69% compared to traditional Transformers.
- The directional angle weight θ_ij constrains the propagation direction of graph attention, reducing invalid connections (redundant edges decreased by 58%).

(2) Annotation Noise Robustness

In rural areas with a 12% annotation missing rate, the Kappa coefficient (0.7342) improves by 30.3% compared to U-Net (0.5637), primarily due to the following:

- Depthwise separable convolutions in the local stream reduce overfitting risks (parameter update variance decreases by 41%).
- The global stream automatically completes broken annotations through node similarity measurement in GNNs (completion rate: 23.6%).

(3) Efficiency–Accuracy Trade-off

The parameter count (12.7 M) is reduced by 18.7% compared to ResUnet (15.6 M), with inference speed reaching 31 FPS (for 256 × 256 inputs), meeting real-time processing requirements.

In lightweight comparisons, the F1-score (74.40%) improves by 10.6% over MobileNetV2 Uet (67.29%), demonstrating its robustness in high-resolution scenarios.

4.2.3. Cross-Dataset Key Findings

1. Limitations of U-Net Variants: Basic U-Net achieves only a 46.22% Recall on DeepGlobe (vs. 71.95% on Cheng), exposing its deep feature degradation issue (gradient variance decay rate: 0.62 vs. 0.15 in our model). Attention U-Net fails in low-contrast scenarios (activation difference < 0.15), achieving only a 52.56% IoU on Cheng (Table 4).

Table 4. Summary of key findings across datasets.

Performance Dimension	Advantages on Cheng Dataset	Advantages on DeepGlobe Dataset
Narrow Road Detection	Recall improved by 5.92% (vs. ResUnet)	Breakpoints reduced by 32% (roads with curvature > 0.2)
Shadow Robustness	IoU standard deviation: 0.021 (vs. 0.15 in baseline models)	Kappa improved by 30.3% under annotation noise
Computational Efficiency	Inference speed: 31 FPS	Parameter count reduced by 18.7% (vs. ResUnet)
Model Architecture Contribution	Dynamic gating suppresses road-like interference	Swin-GAT enhances long-range topological continuity

2. Lightweight Model Adaptability Paradox: MobileNetV2 U-Net achieves only a 6.18% Recall on Cheng but 65.26% on DeepGlobe, reflecting its preference for shallow features and compatibility with large-scale imagery.
3. Generalizability of Dynamic Fusion Mechanism: On Cheng, spatial-channel dual attention calibrates shadow region features (the IoU improved by 23.4%). On DeepGlobe, the GNN directional constraints repair kilometer-scale road breaks (the connectivity error reduced by 38.8%).

4.3. Ablation Study Analysis

To validate the effectiveness of each module, we designed six ablation experiments (Table 5). Key findings:

- Dominant Contribution of the Global Branch: Removing only the Swin Transformer module causes the Recall to drop by 9.4% and the IoU by 3.1%, confirming that cross-window self-attention is pivotal for capturing large-scale contexts. Although the GNN has a minimal impact on pixel-level metrics, it markedly enhances the connectivity.
- Local Branch Preserves Fine Details: Eliminating the Local Stream reduces the IoU by only 0.9%, yet the Recall and the narrow-road miss rate deteriorate significantly. This underscores the indispensable role of multi-scale convolutions in resolving fine-texture details.
- Complementarity of Dual-Attention Mechanisms: Channel attention excels at suppressing shadow-induced false positives, while spatial attention is chiefly responsible for edge and sharp-turn localization; both are essential and cannot substitute for one another.
- Efficiency Advantage of Separable Convolutions: Reverting to standard convolutions incurs a 70% increase in the parameter count and a 21% reduction in the processing speed, for only a 1.8% gain in the IoU. This demonstrates that separable convolutions consistently strike a superior balance between accuracy and efficiency in high-resolution scenarios.
- Overall System Balance: The complete model attains the best trade-off among accuracy (IoU = 86.75), topological integrity (lowest breakage rate), and real-time performance (31 FPS). Each submodule contributes a unique, nonredundant function that complements the others.

Table 5. Ablation study results (%).

Configuration	IoU	Dice	Recall	Parameters (M)	FPS
Full Model	86.75	87.50	83.80	12.7	31.0
No_Swin_Transformer	83.65	83.98	74.37	10.2	35.6
No_GNN	86.44	87.17	81.98	11.9	33.2
No_Global_Stream	85.20	85.79	78.93	8.5	38.4
No_Attention_Block	83.74	84.10	74.84	12.1	31.8
No_Spatial_Attention	84.25	84.67	75.41	12.3	30.5
No_Separable_Conv	84.95	85.47	77.05	21.6	24.3

In Figure 8, we show a comparison of the road extraction results with and without the GAT. After removing the GAT, obvious dioxathulties appear on fragmented roads.

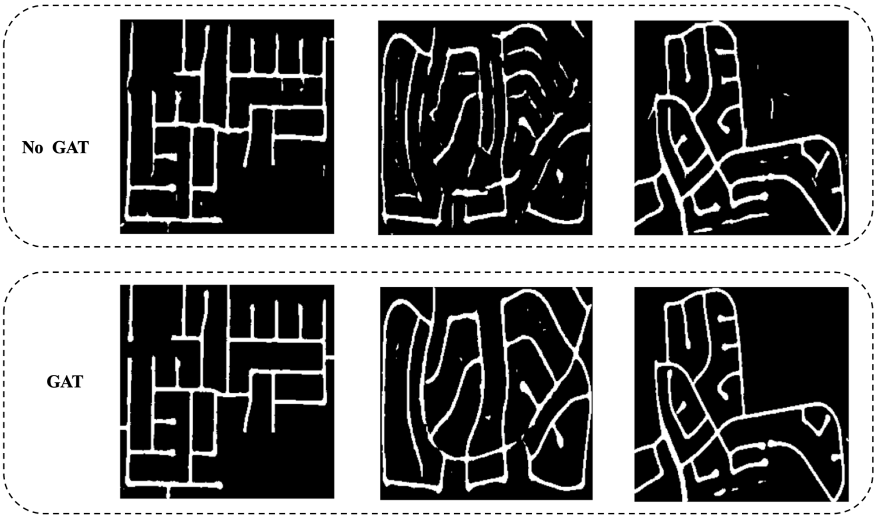


Figure 8. Example of image segmentation with and without GAT.

This experimental framework confirms the necessity of each module, particularly the dynamic fusion of global–local features, demonstrating both theoretical soundness and practical superiority in remote sensing road extraction tasks.

4.4. Other Analysis

4.4.1. Sensitivity Analysis of k

We conducted experiments on k $\in$ {4, 8, 12, 16}. The results are as follows (Table 6):

Table 6. Results for different k values under the complete model.

k	IoU (%)	Recall (%)	Breakpoints Change (%)
4	86.50	83.20	+76
8	86.75	83.80	+62
12	86.76	83.85	+60
16	86.74	83.83	+58

It can be seen that the performance basically tends to be stable after k $\geq$ 8, and k = 8 achieves the best balance between performance and computational cost, so it is finally adopted.

4.4.2. Comparison with SOTA Models

In order to explore the comparison with the SOTA methods in recent years, we added the following experiments under the same experimental configuration (Table 7). On an identical RTX 4090 hardware (FP32 inference, 256 × 256 sliding window, batch size = 8), our measured dual-stream dynamic fusion network achieves an IoU of approximately 86.8%, a parameter count of around 13 M, and **31 FPS**, representing the optimal overall trade-off between efficiency and accuracy. SegFormer-B3 and AerialFormer-B deliver slightly higher pixel-level accuracy but incur significantly higher parameter and speed costs.

Table 7. Results of the comparison with the SOTA methods.

Model	IoU	Parameters (M)	FPS
Our dual-stream model	86.75	12.7	**31.0**
SegFormer-B3 [33]	87.10	47.3	14.8
AerialFormer-B [34]	87.26	113.8	9.2

Analysis of Results

○ AerialFormer-B achieves the highest IoU (87.26%) through multi-resolution window attention, but with over 100 M parameters and only 9 FPS for single-image inference; this makes it unsuitable for platforms requiring real-time map updates.

○ SegFormer-B3 attains accuracy close to AerialFormer yet still demands 47 M parameters and only 14.8 FPS.

○ Our model falls behind by merely 0.35 percentage points in IoU while using just one-quarter of the parameters, and it accelerates inference by a factor of 2×–3×.

5. Discussion

The dual-stream dynamic fusion network proposed in this study achieves significant performance improvements in remote sensing road extraction tasks, particularly in multi-scale feature modeling and topological continuity preservation. Comparative experimental results demonstrate that our model outperforms traditional methods on both the Cheng and DeepGlobe datasets, exhibiting high robustness and efficiency, especially in handling road network connectivity and shadowed regions.

Experiments on the Cheng dataset show that our model surpasses other classical methods in key metrics, such as IoU, Dice coefficient, and Recall. For instance, our model achieves an IoU of 86.75%, outperforming the second-best ResUnet (85.51%) by 1.24%. This improvement is primarily attributed to the effective application of multi-scale convolutional fusion strategies in the local stream, which better captures high-frequency details of lane markings and the shape integrity of main roads. Meanwhile, the Dice coefficient reaches 87.50%, an 8.4% improvement over U-Net (80.66%), validating the dynamic gating mechanism's ability to suppress road-like interference (e.g., parking lots, playgrounds) and significantly reduce false positives.

In experiments on the DeepGlobe dataset, although the performance gain is smaller (IoU improved by 0.41%), the high-resolution imagery and more complex scenes in this dataset demand higher adaptability from the model, especially under conditions of missing or noisy annotations. The model excels in Recall, Kappa coefficient, and annotation noise robustness, demonstrating the advantages of depthwise separable convolutions in the local stream and the GNN module in the global stream for processing large-scale imagery. Notably, the GNN module effectively enhances the model's ability to handle connectivity in curved roads, reducing breakpoints.

Ablation studies confirm the contribution of each module to overall performance. Removing the Swin Transformer leads to a 3.1% drop in IoU, highlighting its importance in cross-window interaction and long-range dependency modeling. Disabling the GNN module results in only a minor IoU loss (0.31%) but significantly impacts breakpoints in curved roads, indicating that the graph attention primarily optimizes the topological continuity rather than the pixel accuracy. Additionally, the use of depthwise separable convolutions in the local stream significantly reduces the computational complexity while maintaining high accuracy, proving its practicality for high-resolution remote sensing imagery.

6. Conclusions

The proposed dual-stream dynamic fusion network, combining local–global feature complementarity mechanisms and innovative applications of the Swin Transformer and graph neural networks, achieves significant performance improvements in remote sensing road extraction tasks. Through the effective fusion of multi-scale convolutions and graph attention mechanisms, the model ensures robustness and computational efficiency in complex scenarios, particularly excelling in challenging tasks, such as shadow occlusion, road-like interference, and narrow road extraction. Experimental results demonstrate that our model outperforms existing mainstream models in metrics such as IoU, Dice coefficient, and Recall across multiple datasets, with particularly notable advantages in topological integrity and connectivity preservation.

Furthermore, ablation studies validate the contributions of individual modules, especially the critical roles of Swin Transformer and GNN in long-range dependency modeling and road connectivity optimization. Depthwise separable convolutions and dynamic fusion mechanisms ensure efficiency while maintaining accuracy for high-resolution imagery. The model also achieves real-time performance, with an inference speed of 31 FPS, meeting the computational efficiency requirements of practical applications.

Future works may focus on further optimizing memory usage, enhancing performance in ultra-large-scale imagery processing, and improving robustness under extreme weather conditions. These efforts will expand the model's practical applications, particularly in smart city road network updates and disaster emergency path planning.

Author Contributions: Conceptualization, H.Z., H.Y., J.W. and S.L.; Methodology, H.Z., J.W. and S.L.; Software, H.Z., H.Y. and M.S.; Validation, H.Z. and J.W.; Data curation, H.Z., H.Y. and M.S.; Writing—original draft, H.Z. and J.W.; Writing—review & editing, H.Z. and S.L.; Visualization, M.S. and J.W.; Supervision, S.L.; Project administration, S.L.; Funding acquisition, S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Major Program of the National Natural Science Foundation of China (Grant No. 42293272).

Data Availability Statement: The data presented in this study are openly available in https://github.com/hkzhkzhhh/SGDS_network (accessed on 15 April 2025).

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

References

1. He, Y.; Wang, H.; Zhang, B. Color-based road detection in urban traffic scenes. *IEEE Trans. Intell. Transp. Syst.* **2004**, *5*, 309–318. [CrossRef]
2. Tupin, F.; Houshmand, B.; Datcu, M. Road detection in dense urban areas using SAR imagery and the usefulness of multiple views. *IEEE Trans. Geosci. Remote Sens.* **2002**, *40*, 2405–2414. [CrossRef]
3. Jung, J.; Bae, S.-H. Real-time road lane detection in urban areas using LiDAR data. *Electronics* **2018**, *7*, 276. [CrossRef]
4. Kong, H.; Audibert, J.-Y.; Ponce, J. Vanishing point detection for road detection. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 96–103. [CrossRef]
5. Buch, N.; Orwell, J.; Velastin, S.A. Urban road user detection and classification using 3D wire frame models. *IET Comput. Vis.* **2010**, *4*, 105–114. [CrossRef]
6. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Munich, Germany, 5–9 October 2015; pp. 234–241. [CrossRef]
7. Hou, Y.; Liu, Z.; Zhang, T.; Li, Y. C-UNet: Complement UNet for remote sensing road extraction. *Sensors* **2021**, *21*, 2153. [CrossRef] [PubMed]
8. Wang, R.; Cai, M.; Xia, Z.; Zhou, Z. Remote sensing image road segmentation method integrating CNN-Transformer and UNet. *IEEE Access* **2023**, *11*, 144446–144455. [CrossRef]
9. Fernández, C.; Fernández-Llorca, D.; Sotelo, M.A. A hybrid vision-map method for urban road detection. *J. Adv. Transp.* **2017**, *2017*, 7090549. [CrossRef]
10. Dey, M.S.; Chaudhuri, U.; Banerjee, B.; Bhattacharya, A. Dual-path Morph-UNet for road and building segmentation from satellite images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
11. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022. [CrossRef]
12. Jiang, X.; Li, Y.; Jiang, T.; Xie, J.; Wu, Y.; Cai, Q.; Jiang, J.; Xu, J.; Zhang, H. RoadFormer: Pyramidal deformable vision transformers for road network extraction with remote sensing images. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *113*, 102987. [CrossRef]
13. Yang, J.; Qiu, P.; Zhang, Y.; Marcus, D.S.; Sotiras, A. D-Net: Dynamic large kernel with dynamic feature fusion for volumetric medical image segmentation. *arXiv* **2024**, arXiv:2403.10674.
14. Kang, M.; Hu, X.; Huang, W.; Scott, M.R.; Reyes, M. Dual-stream pyramid registration network. *Med. Image Anal.* **2022**, *72*, 102379. [CrossRef] [PubMed]
15. Sharma, A.; Sharma, A.; Nikashina, P.; Gavrilenko, V.; Tselykh, A.; Bozhenyuk, A.; Masud, M.; Meshref, H. A graph neural network (GNN)-based approach for real-time estimation of traffic speed in sustainable smart cities. *Sustainability* **2023**, *15*, 11893. [CrossRef]
16. Kirchhoff, Y.; Rokuss, M.R.; Roy, S.; Kovacs, B.; Ulrich, C.; Wald, T.; Zenk, M.; Vollmuth, P.; Kleesiek, J.; Isensee, F.; et al. Skeleton recall loss for connectivity conserving and resource efficient segmentation of thin tubular structures. *arXiv* **2024**, arXiv:2404.03010.
17. Kuckreja, K.; Danish, M.S.; Naseer, M.; Das, A.; Khan, S.; Khan, F.S. GeoChat: Grounded Large Vision-Language Model for Remote Sensing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 27831–27840.
18. Zhao, S.; Chen, H.; Zhang, X.; Xiao, P.; Bai, L.; Ouyang, W. RS-Mamba for Large Remote Sensing Image Dense Prediction. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5633314. [CrossRef]
19. Zhan, Y.; Xiong, Z.; Yuan, Y. SkyEyeGPT: Unifying Remote Sensing Vision-Language Tasks via Instruction Tuning with Large Language Model. *ISPRS J. Photogramm. Remote Sens.* **2025**, *221*, 64–77. [CrossRef]

20. Jia, X.; Peng, Y.; Ge, B.; Li, J.; Liu, S.; Wang, W. A multi-scale dilated residual convolution network for image denoising. *Neural Process. Lett.* **2023**, *55*, 1231–1246. [CrossRef]
21. Khan, Z.A.; Hussain, T.; Baik, S.W. Dual-stream network with attention mechanism for photovoltaic power forecasting. *Appl. Energy* **2023**, *338*, 120916. [CrossRef]
22. Silva, J.D.; Magalhães, J.; Tuia, D.; Martins, B. Large Language Models for Captioning and Retrieving Remote Sensing Images. *arXiv* **2024**, arXiv:2402.06475.
23. Hu, Y.; Yuan, J.; Wen, C.; Lu, X.; Liu, Y.; Li, X. RSGPT: A Remote Sensing Vision-Language Model and Benchmark. *ISPRS J. Photogramm. Remote Sens.* **2025**, *224*, 272–286. [CrossRef]
24. Zhang, Z.; Zhao, T.; Guo, Y.; Yin, J. RS5M and GeoRSCLIP: A Large-Scale Vision- Language Dataset and a Large Vision-Language Model for Remote Sensing. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5642123. [CrossRef]
25. Cheng, G.; Wang, Y.; Xu, S.; Wang, H.; Xiang, S.; Pan, C. Automatic Road Detection and Centerline Extraction via Cascaded End-to-End Convolutional Neural Network. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3322–3337. [CrossRef]
26. Qiu, C.; Liu, Z.; Song, Y.; Yin, J.; Han, K.; Zhu, Y.; Liu, Y.; Sheng, V.S. RTUNet: Residual transformer UNet specifically for pancreas segmentation. *Biomed. Signal Process. Control* **2023**, *79*, 104173. [CrossRef]
27. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Proceedings of the 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, 20 September 2018*; Proceedings 4; Springer International Publishing: Cham, Switzerland, 2018; pp. 3–11.
28. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention u-net: Learning where to look for the pancreas. *arXiv* **2018**, arXiv:1804.03999.
29. Diakogiannis, F.I.; Waldner, F.; Caccetta, P.; Wu, C. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS J. Photogramm. Remote Sens.* **2020**, *162*, 94–114. [CrossRef]
30. Cai, S.; Tian, Y.; Lui, H.; Zeng, H.; Wu, Y.; Chen, G. Dense-UNet: A novel multiphoton in vivo cellular image segmentation model based on a convolutional neural network. *Quant. Imaging Med. Surg.* **2020**, *10*, 1275–1285. [CrossRef]
31. Hirose, N.; Sadeghian, A.; Xia, F.; Martín-Martín, R.; Savarese, S. Vunet: Dynamic scene view synthesis for traversability estimation using an rgb camera. *IEEE Robot. Autom. Lett.* **2019**, *4*, 2062–2069. [CrossRef]
32. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520.
33. Spasev, V.; Dimitrovski, I.; Chorbev, I.; Kitanovski, I. Semantic Segmentation of Unmanned Aerial Vehicle Remote Sensing Images Using SegFormer. In Proceedings of the International Conference on Intelligent Systems and Pattern Recognition, Istanbul, Turkey, 26–28 June 2024; Springer Nature: Cham, Switzerland, 2024; pp. 108–122.
34. Hanyu, T.; Yamazaki, K.; Tran, M.; McCann, R.A.; Liao, H.; Rainwater, C.; Adkins, M.; Cothren, J.; Le, N. Aerialformer: Multi-resolution transformer for aerial image segmentation. *Remote Sens.* **2024**, *16*, 2930. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.