Signal, Image and Video Processing (2024) 18:2403–2412
https://doi.org/10.1007/s11760-023-02916-1

ORIGINAL PAPER

RoadTransNet: advancing remote sensing road extraction through multi-scale features and contextual information

K. Madhan Kumar[1]

Received: 16 October 2023 / Revised: 20 November 2023 / Accepted: 21 November 2023 / Published online: 27 December 2023
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2023

Abstract
Road extraction is a crucial task that requires high-resolution remote sensing images with wide-ranging applications in urban planning, navigation, and autonomous vehicles. However, this task is challenged by complex road structures and the need to capture long-range dependencies. RoadTransNet is a new road extraction architecture that aims to solve these problems that making the power of the Swin Transformer and Feature Pyramid Network (FPN) while introducing Transformer-like attention mechanisms. RoadTransNet combines a robust convolutional backbone, inspired by the Swin Transformer, with an FPN to capture multi-scale features effectively. The Transformer-like attention mechanisms, including multi-head self-attention and cross-attention, enable the network to represent context information on a local and global scale, ensuring accurate road extraction. The skip connections facilitate gradient flow, preserving fine details, and decoding layers transform extracted features into precise road predictions. Our experiments are conducted using the RoadTransNet, which is subject to rigorous assessment on the following datasets: the DeepGlobe road extraction challenge Dataset and the CHN6-cUG roads dataset. The outcomes indicate its superior performance in achieving high-level metrics of precision and recall, as well as achieving high F1 scores and IoU. The comparative evaluations performed against traditional methods showcase RoadTransNet's ability to capture complex road structures and long-range dependencies. The RoadTransNet stands as a comprehensive solution for the extraction of roads in high-resolution remote sensing images, offering promising opportunities for improving urban planning, navigation systems, and autonomous vehicle technologies. Its success lies in the synergy of convolutional and transformer-based architectures, paving the way for advanced remote sensing applications in smart cities and others.

Keywords RoadTransNet · Swin transformer · Feature pyramid network · Self-attention · Cross-attention · Skip connections · Decoder · Road extraction

1 Introduction

Remote sensing road extraction acts a fundamental task in a variety of essential applications, ranging from urban planning and navigation systems to the development of autonomous vehicles [1]. The provision of accurate and timely road network data is essential for the management of urban infrastructure, traffic analysis, disaster response, and emerging technologies such as self-driving cars. With the advent of high-resolution satellite and aerial imagery, the potential for automating road extraction processes has grown significantly. However, extracting road networks from complex and diverse environments remains a formidable challenge [2]. Furthermore, urban environments often encompass long-range dependencies among road segments, which cannot be effectively captured by traditional computer vision methods [3].

Traditional road extraction techniques, including texture analysis and mathematical morphology, have been time-consuming and error-prone, as they heavily rely on hand-crafted features and human operator intervention [4]. One of the foremost challenges is the intrinsic complexity of urban environments, which often leads to the misclassification of roads due to their visual similarity with other urban features. In areas characterized by complex urban settings, it becomes essential to extract roads with high precision while minimizing false positives.

Additionally, remote sensing (RS) imagery tends to lack effective capture of long-distance dependencies using

✉ K. Madhan Kumar
madhankn@gmail.com

[1] Department of Electronics and Communication Engineering, PET Engineering College, Valliyur, Tamilnadu, India

DCNN-based models [5]. Existing deep learning architectures, while adept at local feature extraction, struggle to capture the global context required to disambiguate roads from other objects accurately. Motivated by the pressing need for more robust road extraction methods capable of addressing the challenges presented by complex urban environments, this research introduces a novel approach. We propose the integration of a Transformer-like attention mechanism into the road extraction pipeline, aiming to improve the model's capacity to detect long-distance dependencies and contextually relevant data. The Transformer architecture, initially designed for natural language processing, has exhibited remarkable success in various computer vision tasks by modeling dependencies among different positions in an image or sequence. Transformers rely on the self-attention mechanism as a fundamental component, which enables the capture of global context information by assigning attention weights to all positions within an input sequence. We hypothesize that by adapting Transformer-like attention to road extraction, we significantly improve the model's ability to identify roads amidst complex urban elements.

Our research on RoadTransNet makes several significant contributions to the field of road extraction:

- *Innovative architecture:* RoadTransNet is a novel architecture that effectively addresses the challenges of complex road structures and long-range dependencies. By combining elements from the Swin Transformer, Feature Pyramid Network (FPN), and Transformer-like attention mechanisms, a powerful and versatile road extraction model is developed.
- *Effective feature extraction:* RoadTransNet leverages a Swin Transformer-based convolutional backbone, pretrained on ImageNet, to extract rich visual features. This backbone provides a strong foundation for the model and is designed to capture representations at both the low and high levels of input RS images.
- *Contextual information:* Capturing of local and global contextual information is enabled by the use of multi-head self-attention and cross-attention mechanisms. This helps to bridge between the finer details and the large-scale dependencies.
- *Multi-scale features:* The Feature Pyramid Network (FPN) is integrated to enhance feature representation at multiple scales. FPN's lateral connections and top-down pathways enable the model to detect road structures of varying widths, high-resolution imagery requires a crucial capability to extract roads.

The leftover portions of this research work are Sect. 2 presents literature relating that focuses on extracting and refining road features from RS imagery for planning development schemes. Section 3 outlines the RoadTransNet model proposed to improve road feature extraction capability is discussed. In Sect. 4, the dataset used for experimentation and the preprocessing steps applied are illustrated. In Sect. 5, the results of the experiment, which were conducted using both qualitative and quantitative measures, are displayed. The paper concludes in Sect. 6 with recommendations for future research.

2 Related work

Multiple approaches designed with the aim of increasing road extraction performance are summarized as follows: attention-based cascaded network was introduced by Li et al. [6]. Li et al. [7] illustrated a RemainNet model for extracting roads from RS images. Images from the two road extraction datasets were used as test, validation, and training sets. This method achieved only a 0.64 IoU value which requires further improvements to increase segmentation accuracy.

Luo et al. [8] demonstrated the use of RS images to create a bidirection transfer network (BDTNet) for extracting roads. The BDTNet model yielded an IoU result of 67.21% when evaluated using the DeepGlobe dataset. TransRoadNet model was introduced by Yang et al. [9] for extracting roads utilizing RS images. This model achieved better road extraction accuracy but has poor generalization ability. HU et al. [10] suggested a multiscale deformable transfer network (MDTNet). This method has achieved high IoU outcomes of about 71.19% for DeeoGlobal road dataset. Wang et al. [11] have developed a special type of neural network called the dual-decoder-U-Net (DDU-Net) to make small-scale road extraction more reliable and accurate when multiple roads of different sizes were combined. Yan et al. [12] have introduced a new regularized surface extraction model using a graph-based neural network. Meanwhile, getting road surface maps that were regular and well-defined was tough, and a lot of manual labor was usually needed. Chandra et al. [13] have demonstrated a technique for the detection of roads from HRSi to focus on comprehending the cognitive processes, knowledge, and reasoning utilized by the analyst when performing the task.

Abdollahi et al. [14] have introduced a new type of deep learning network called a VNet model that could create high-quality road segmentation maps. Wei et al. [15] have introduced a multistage framework to extract both road surface as well as road centerline at the same time. However, obtaining roads from high-resolution RMS images was still challenging due to tree and building shadowing, road discrimination, and intricate backgrounds. Luo et al. [16] have designed AD-RoadNet that could help with decoding roads. To assess the designed AD-RoadNet, additional information such as ablation analysis, inference size matter, and the labels was of poor quality, and there were lots of different images

that showed different levels of quality. Yin et al. [17] have introduced a road extraction network which was referred to as C2S-RoadNet. This model was capable of establishing long-distance connections and making full use of global data, and it was able to extract road information more effectively. Yang et al. [18] introduced the SSEANet a semi-permeable edge-aware network that enables RS of image segmentation.

The Roadformer by Jiang et al. [19] suffers to extracts road features because of inherent difficulty to maintain extraction at high level at constant time. Christophe et al. [20] have introduced a robust geometrical approach to provide an initial extraction level. The algorithm was highly efficient and had a limited number of parameters Lan et al. [29] have introduced an automated road extraction module utilizing global context aware dilated convolution network. The implementation of residual dilated layers has enlarged the receptive field and aids in learning extra discriminative features. To extract and congregate multidimensional global contextual information, a pyramid pooling network was utilized. Their integration has entitled the stronger feature representation capability of framework. It performs well on small scale dataset while not examined using large scale dataset.

2.1 Limitations of existing methods

While deep learning-based methods have demonstrated substantial progress in road extraction, they continue to grapple with several limitations, especially when dealing with complex urban environments: *(i) Intra-class Differences and Interclass Similarities*: Roads come in various forms, including urban roads, rural roads, highways, and railways, each with distinct intraclass differences. Furthermore, urban roads often share high interclass similarities with other urban features such as buildings and pavements, (ii) *Failure in Complex Scenes*: Many state-of-the-art models that excel in relatively simple scenes struggle when applied to complex urban environments. The intricate urban layout and the presence of diverse objects challenge the ability of these models to distinguish roads from other features. (iii) *Limited Long-Range Dependency Modeling*: current neural network models often fall short of accurately capturing RS imagery long-distance dependencies, (iv) *High computational overhead:* Some deep learning architectures involve significant computational overhead, making them less practical for real-time or large-scale road extraction tasks. In light of these limitations, this research introduces a novel approach that leverages Transformer-like attention mechanisms to address the challenges associated with capturing long-range dependencies and contextual information in road extraction. The proposed "RoadTransNet" architecture aims to revolutionize the field by offering improved precision and generalization across diverse urban environments.

3 Proposed methodology: RoadTransNet

RoadTransNet is a novel road extraction architecture designed to overcome the limitations of existing methods by effectively capturing long-range dependencies and contextual information. At its core, RoadTransNet (shown in Fig. 1) combines a convolutional backbone, a Feature Pyramid Network (FPN), and a Transformer-like attention mechanism. Each of these component's working procedures is detailed in the following sub-sections.

3.1 Convolutional backbone layer

In RoadTransNet, the selection of a robust convolutional backbone is a pivotal decision as it determines ability of network to identify useful features in high-resolution RS images. For this purpose, the Swin Transformer architecture is chosen as our convolutional backbone. Swin Transformer has gained widespread recognition for its effectiveness in capturing hierarchical features, making it an ideal choice for our task of extracting roads.

Swim transformer module: The Swin Transformer architecture is characterized by its unique design, which combines the strengths of both convolutional and self-attention mechanisms. The Convolutional Backbone consists of multiple Swin Transformer layers stacked on top of each other. This hierarchical feature extraction is essential for detecting road structures at different scales and complexities.

Model pretraining: The convolutional backbone is trained on ImageNet, prior to fine-tuning for road extraction. This pretraining step helps the network learn rich visual features from a diverse range of images, which are then leveraged for the task of road extraction.

Each Swin Transformer layer within the Convolutional Backbone is composed of the following key components: (i) *Patch Embedding:* The input image is subdivided into non-intersecting segments, and each segment is inserted into a sub-dimensional space. This allows the network to focus on localized information within each patch. (ii) *Multi-Head Self-Attention:* In Swin Transformer, it allows each patch to assist to other patches both locally and globally. This captures important contextual information and long-range dependencies within the image. (iii) *Feedforward Neural Networks:* After attention mechanisms, the features are passed through feedforward neural networks (FFNs) within each patch. These networks perform non-linear transformations and capture complex relationships between features. (iv) *Residual Connections:* To facilitate gradient flow and prevent vanishing gradients, residual connections are added around each layer. These connections allow gradients to flow smoothly during training, enabling the network to obtain low-level as well as high-level features efficiently.

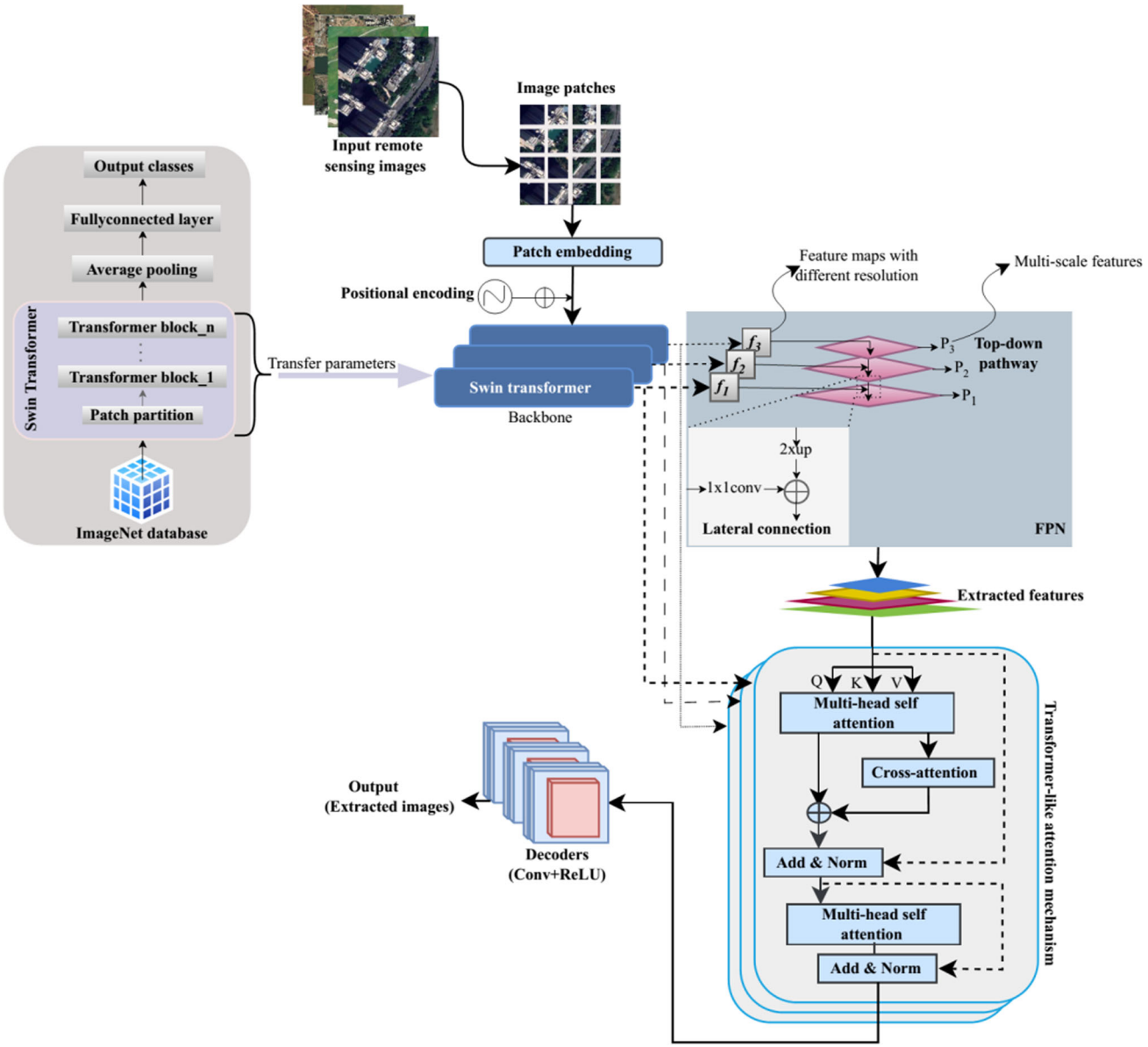


Fig. 1 Overall architecture of the proposed road extraction framework

These operations, combined with the hierarchical organization of the Swin Transformers layers, lead to the extraction of high-quality features from the input images.

3.2 Feature pyramid network

FPN is integrated after the Convolutional Backbone to enhance feature representation at multiple scales. This is particularly important for addressing the challenge posed by multiscale road the structures. The top-down pathways and lateral connections make up the FPN. These connections enable the extraction of features at different spatial resolutions, ensuring that the network effectively detects both narrow and wide roads.

Lateral Connections: FPN includes lateral connections that connect feature maps from different levels of the Convolutional Backbone. For each lateral connection, consider the feature map at level i as P_i and the feature map from the layer above $(i + 1)$ as $P(i + 1)$. Lateral connections are achieved through 1×1 convolutions, and the equation for lateral connections can be expressed as:

$$P_i' = 1 \times 1 Conv\left(P_{(i+1)}\right) \tag{1}$$

here, P_i' represents the feature map from lateral connections, which is the result of applying a 1×1 convolution to $P_{(i+1)}$.

Top-down pathways: In addition to lateral connections, FPN incorporates top-down pathways that involve upsampling and merging feature maps from higher resolutions to lower resolutions. It results in feature pyramids with rich information at multiple scales. It is mathematically represented as,

$$P'_{(i-1)} = \text{Upsample}\left(P'_i\right) + P_i \tag{2}$$

In this equation, $P'_{(i-1)}$ represents the upsampled feature map from the layer below $(i-1)$. Upsample() denotes the upsampling operation to match the spatial dimensions, and + represents element-wise addition.

3.3 Transformer like attention mechanism

RoadTransNet's transformer-like attention mechanism helps capture global and local context. The attention mechanism operates in self-attention and cross-attention manner.

3.3.1 Self-attention

Self attention is an essential component of the transformer architecture that enables each pixel in the feature map to weigh its importance concerning other pixels. In the context of road extraction, self-attention allows RoadTransNet to concentrate on specific road segments while leaving out irrelevant background data. For a given feature map $X \in \Re^{(H\times W\times C)}$, where H,W, and C represent the height, width, and channel numbers, respectively, self-attention computes a weighted combination of the feature vectors at different positions:

$$Y = \text{Soft max}\left(\frac{QK^T}{\sqrt{d_k}}\right).V \tag{3}$$

where, Q, V and K are linearly projected queries, values and keys derived from input feature map X; Soft max is the softmax function that normalizes the attention scores; d_k implies key vector size.

This operation allows RoadTransNet to attend to relevant spatial locations while suppressing noise and irrelevant details.

3.3.2 Cross-attention

In addition to self-attention, RoadTransNet incorporates cross-attention to capture global contextual information. Unlike self-attention, cross-attention uses two distinct input sequences, which means that, queries from input sequence x, and values and keys from input sequence y with similar embedding sizes. Cross-attention considers information from other image regions, enabling it to understand the relationships between roads and surrounding features. Mathematically, cross-attention is expressed as:

$$Y = \text{Soft max}\left(\frac{Q_x K_y^T}{\sqrt{d_k}}\right).V_y \tag{4}$$

where K_y and V_y are linearly projected keys and values derived from a different portion of the input feature map y, and Q_x is the linearly projected query from input sequence x. By combining self-attention and cross-attention mechanisms, RoadTransNet achieves a balance between local and global information, allowing it to effectively capture road context in complex scenes.

3.3.3 Positional encodings

To enable the transformer-like attention mechanism to consider the spatial relationships between pixels, the Addition of positional encodings to the input feature map. Position encoding is a data structure that describes the absolute and approximate positions of pixels in an image.

3.4 Skip connections and decoding layers

To enhance the training process and ensure that RoadTransNet effectively captures both high-level semantics and low-level details, the skip connections and decoding layers are introduced into the architecture.

3.4.1 Skip connections

Skip connections allow the network to recover low-level features from the Convolutional Backbone. Mathematically, skip connections are realized through element-wise addition, as follows:

$$X_{\text{skip}} = X_{\text{Conv\_ Backbone}} + X_{\text{TA}} \tag{5}$$

where X_{skip} represents the feature maps obtained through skip connections, $X_{\text{Conv\_Backbone}}$ denotes the feature maps from the convolutional backbone and X_{TA} signifies the high level representations generated by the transformer like attention mechanism.

3.4.2 Decoding layers

The decoding process involves a series of convolutional and activation layers that progressively refine the features. We leverage the contextual understanding learned by the attention mechanism to guide the decoding process and ensure that the final predictions align with the road structures present in the input images. Mathematically, the decoding process is

represented as:

$$P_{\text{road}} = \text{Decoder}(X_{\text{skip}}) \tag{6}$$

where P_{road} represents the pixel-wise road probability map and $\text{Decoder}(X_{\text{skip}})$ denotes the decoding layers responsible for refining the features.

3.5 Activation function and loss function

The activation/transfer function utilized in RoadTransNet is Rectified Linear Unit (ReLU). It introduces non-linearity into the network, which is essential for learning complex road patterns and features. For RoadTransNet training, the loss/error function is the Dice Loss. It measures the similitude between the ground truth and predicted road masks, making it effective for handling imbalanced datasets and encouraging accurate road extraction. Using the following expression, the Dice Loss is computed,

$$\text{Dice Loss} = 1 - \frac{2 \cdot |P \cap G|}{|P| + |G|} \tag{7}$$

where P represents the predicted road mask, G represents the ground truth road mask, $|P \cap G|$ is the intersected mask regions between ground truth and predicted, $|P|$ is predicted mask of the total number in road pixel, and $|G|$ is the total road pixel count in the ground truth mask.

4 Experimental data and preprocessing

In this section, a detailed overview of the datasets utilized in RoadTransNet for assessment and training is displayed, along with the specific data preprocessing steps, augmentation techniques, and the strategy employed for dataset splitting.

4.1 Dataset description

Two primary datasets are employed for training and evaluating RoadTransNet:

DeepGlobe Road Extraction Challenge Dataset [21, 22]: This dataset constitutes a fundamental component of our training and evaluation data. It encompasses high-resolution remotely sensed imagery obtained from areas in Thailand, India, and Indonesia, with the following attributes. The imagery boasts an impressive spatial resolution of 50 cm per pixel, ensuring fine-grained visual details. Each image in the dataset is sized at 1024×1024 pixels, offering an extensive field of view for comprehensive road analysis. In accordance with common practice, we partition the dataset is divided into 1250 testing images and 4976 training images.

CHN6-CUG Roads Dataset [23]: It is another pivotal component of our training and evaluation data and is curated from Google Earth. It comprises images captured in 6 representative cities of China, featuring following characteristics: Image's spatial resolution is 50 cm per pixel, allowing for detailed urban and suburban analysis. Each image is formatted to 512×512 pixels, providing a balanced representation of different urban landscapes. The dataset is thoughtfully divided into two subsets: 3608 training images are reserved and 903 testing images are assigned.

4.2 Dataset preprocessing

To bolster model robustness and enhance training efficiency, a series of data preprocessing and augmentation techniques are judiciously applied: (i) *Resizing:* Feeding two different dataset images with varied resolutions might impact the model's extraction performance. So, resizing is performed on the images that scale the images into the desired range without any information loss. In our work, the images are resized to a pixel dimension of 512×512. (ii) *Normalization:* Input images undergo normalization to ensure they exhibit zero mean and unit variance. This standardization guarantees uniform scaling across the entire dataset. (iii) *Data Augmentation:* Augmentation procedures, including random horizontal flips, random rotations, and random scaling, are judiciously employed. These augmentations serve to enrich the diversity of the training dataset, empowering the model to generalize effectively across varying road structures and environmental conditions.

4.3 Training process

The training process revolves around the fine-tuning of RoadTransNet on the combined DeepGlobe and CHN6-CUG datasets, with meticulous attention to hyperparameters. RoadTransNet is optimized using the Adam optimization algorithm. The choice of Adam is based on rigorous experimentation to ascertain the best results. Using Adam optimizer, the hyperparameters such as learning rate is set to 2×10^{-4}, maximum training epochs to 100 and batch size to 8.

5 Experimental results

Here, we meticulously show the results of simulation conducted with RoadTransNet, offering both quantitative and qualitative insights into its performance. The entire development of this work is based on Pytorch framework v1.11.0, running on Nvidia GeForce 3060 toolkit using intel core i59400f, 12 GB memory, and Windows 10 operating system. Also, RoadTransNet's performance is rigorously evaluated

Table 1 Performance analysis results of DeepGlobe Dataset

Methods	Metrics				
	Precision (%)	Recall (%)	F1-score (%)	IoU (%)	APLS (%)
A	82.72%	85.71%	84.78	82.83	69.73%,
B	80.92%	83.91%	86.87	84.98	75.71%
C	84.18%	76.77%	80.3	78.66	70.07%
D	70.05%	61.17%	61.35	61.09	69.72%
E	88.67%	85.07%	78.28	76.67	65.41%
F	90.75%	87.72%	89.27	87.66	57.92%
G	76.27%	72.09%	76.75	74.98	60.02%
H	76.27%	79.42%	78.16	76.34	70.13%
I	90.05%	92.85%	91.43	87.67	74.67%
J	93.04%	89.30%	91.35	87.67	87.74%
K	84.37%	82.11%	83.22	80.02	75.64%
Proposed	95.72%	95.06	94.86	92.10	92.47

Table 2 Performance analysis results of CHN6-CUG Roads dataset

Methods	Metrics				
	Precision (%)	Recall (%)	F1-score (%)	IoU (%)	APLS (%)
A	67.02	76.04	75.34	60.43	68.36
B	56.47	58.78	67.53	68.02	74.45
C	76.93	75.49	68.43	65.32	68.34
D	68.43	62.07	71.01	63.80	62.79
E	56.89	80.30	77.90	61.27	64.49
F	87.56	56.87	65.93	56.43	56.76
G	63.09	63.46	69.26	52.87	59.54
H	74.39	72.02	86.32	63.76	70.02
I	80.45	85.43	78.63	70.98	73.78
J	72.03	68,05	84.07	82.46	85.03
K	79.35	79.44	79.39	65.93	81.97
Proposed	93.56	92.76	90.76	89.65	90.03

using a range of quantitative metrics: Precision, Recall, F1-Score, and IoU. These metrics provide a comprehensive assessment of the model's accuracy, completeness and spatial alignment with the ground truth road network.

5.1 Quantitative results

This section evaluates the efficiency of the proposed RoadTransNet in resolving RS road extraction tasks using quantitative metrics.

Table 1 shows the Precision, recall, f1-score, IoU, and APLS rate analysis graph of different methods for the DeepGlobe dataset. This comparison has gauged the exactness to extract road features by the proposed RoadTransNet over others. The RoadTransNet has achieved a precision rate of 95.72% which is the highest among the compared 11 other competitive road extraction methods such as RADANet (Method A) [24], TransRoadNet (Method B) [10], BDTNet (Method C) [9], Swin-ResUnet + (Method D) [25], C2S-RoadNet (Method E) [18], CoupleUNet (Method F) [26], TransLinkNet (Method G) [28], RemainNet (Method H) [8], Seg-Road (Method I) [29], RoadFormer (Method J) [20], AD-RoadNet (Method K) [27]. As, recorded in the table, the each measure evaluated achieved high performance rate, illustrating its quality and accuracy of prediction.

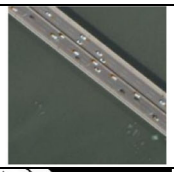
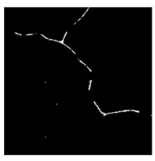
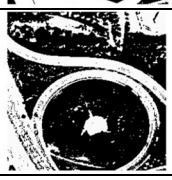
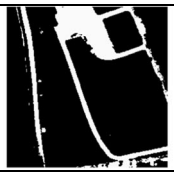
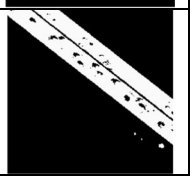
The analysis outcomes of the suggested RoadTransNet for assessing the CHN6-CUG Roads dataset are displayed in Table 2.

The proposed TransRoadNet method gives better results of 93.56% precision, 92.76% Recall, 90.76% f1-score, 89.65% IoU, and 90.03% APLS, in which these values are greater than the existing methods that are compared. Table

Table 3 Ablation experiment

Components							Precision	
Swin Transformer	FPN	Attention mechanism		Positional encoding	Skip connections	Decoding layers	DeepGlobe dataset	CHN6-CUG Roads dataset
		Self-attention	Cross-attention					
✓	X	X	X	X	X	X	0.561	0.532
✓	X	X	X	X	X	✓	0.589	0.563
✓	✓	X	X	X	X	X	0.629	0.602
✓	✓	X	X	X	X	✓	0.682	0.654
✓	✓	✓	X	X	X	X	0.750	0.739
✓	✓	✓	X	X	X	✓	0.785	0.732
✓	✓	X	✓	X	X	X	0.763	0.751
✓	✓	X	✓	X	X	✓	0.785	0.773
✓	✓	✓	✓	X	X	X	0.805	0.796
✓	✓	✓	✓	X	X	✓	0.846	0.836
✓	✓	✓	✓	✓	X	X	0.837	0.816
✓	✓	✓	✓	✓	X	✓	0.865	0.852
✓	✓	✓	✓	✓	✓	X	0.916	0.894
✓	✓	✓	✓	✓	✓	✓	0.957	0.935

Table 4 Visualized road extraction results of CHN6-CUG Roads dataset using RoadTransNet

Images					
Ground truths					
Output images					

3 provides the ablation experiment of our RoadTransNet framework. The result shows that the inclusion of distinct components to perform specific operations has highly enhanced results of extracting roads from background of RS samples.

5.2 Qualitative results

Sample Image Visualizations: We present sample images from the testing dataset, juxtaposing the original remote sensing imagery which corresponds to field truth road masks as well as RoadTransNet's predictions. These visualizations vividly demonstrate the model's effectiveness in capturing road structures, regardless of their complexity or scale.

The road extraction results for CHN6-CUG Roads dataset for various methods are shown in Table 4. Rural roads comprise the majority of the satellite images included in the DeepGlobe dataset, displaying similar texture to the background and covering vegetation areas and forest lands with varied shadow lengths. However, with the integrated components of the Swin Transformer, FPN, and attention

mechanisms, the proposed RoadTransNet extracts or segments complex spatial information from the RS imagery data with high degree of accuracy. In comparison with DeepGlobe dataset, the images in CHN6-CUG dataset consist of more intricate intersections, high spectral reflectance, elevated roads, blocked roads, heavily trafficked roads, railways, narrow paths, etc. The typical extraction techniques face high level of difficulty in extracting road scenes from such complex scenes, resulting in loss of significant features. But, our proposed RoadTransNet model with its ability to extract contextual information extracts more intricate road networks precisely, showing closer results to ground truths. It is visually displayed that the proposed RoadTransNet extracts road features precisely than other compared networks.

6 Conclusion

Extracting roads from RS high-resolution images is one of the most fundamental tasks with significant implications for urban planning, navigation, and autonomous vehicles. Complex road structures and the necessity to capture long-range dependencies have posed formidable challenges to existing methods. In response, we introduced RoadTransNet, an innovative architecture that combines the strengths of the Swin Transformer, Feature Pyramid Network (FPN), and Transformer-like attention mechanisms to address these challenges effectively. The cornerstone of RoadTransNet is its convolutional backbone, inspired by the Swin Transformer architecture. The DeepGlobe Road Extraction Challenge Dataset and CHN6-CUG Roads Datasets were used for the extensive experiments that demonstrated RoadTransNet's outstanding performance. The model consistently achieved high precision, the recall, the F1-score, and the Intersection over Union (IoU) metrics, surpassing traditional methods. RoadTransNet's ability to capture complex road structures and long-range dependencies positions it as a formidable solution in the area of image analysis from remote sensing. Future research will focus on addressing computational intensity to enhance the network's performance and hyperparameter tuning problems.

Authors' contributions KMK agreed on the content of the study. KMK collected all the data for analysis. KMK agreed on the methodology. KMK completed the analysis based on agreed steps. Results and conclusions are discussed and written together. The author read and approved the final manuscript.

Funding Not applicable.

Availability of data and material The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethics approval This article does not contain any studies with human participants.

Human and animal rights This article does not contain any studies with human or animal subjects performed by any of the authors.

Informed consent Informed consent was obtained from all individual participants included in the study.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

1. Yi, F., Te, R., Zhao, Y., Xu, G.: EUNetMTL: multitask joint learning for road extraction from high-resolution RS images. Remote Sensing Letters. **13**(3), 258–268 (2022)
2. Abdollahi, A., Pradhan, B., Alamri, A.: SC-RoadDeepNet: A new shape and connectivity-preserving road extraction deep learning-based network from remote sensing data. IEEE Trans. Geosci. Remote Sens. **60**, 1–15 (2022)
3. Chen, W., Zhou, G., Liu, Z., Li, X., Zheng, X., Wang, L.: NIGAN: A framework for mountain road extraction integrating remote sensing road-scene neighborhood probability enhancements and improved conditional generative adversarial network. IEEE Trans. Geosci. Remote Sens. **60**, 1–15 (2022). https://doi.org/10.1109/TGRS.2022.3188908
4. Zhang, Z., Sun, X., Liu, Y.: GMR-Net: road-extraction network based on fusion of local and global information. Remote Sensing. **14**(21), 5476 (2022)
5. Jie, Y., He, H., Xing, K., Yue, A., Tan, W., Yue, C., Jiang, C., Chen, X.: MECA-net: a multiscale feature encoding and long-range context-aware network for road extraction from remote sensing images. Remote Sensing. **14**(21), 5342 (2022)
6. Li, S., Liao, C., Ding, Y., Hu, H., Jia, Y., Chen, M., Xu, B., Ge, X., Liu, T., Wu, D.: Cascaded residual attention enhanced road extraction from remote sensing images. ISPRS Int. J. Geo Inf. **11**(1), 9 (2022)
7. Li, Z., Chen, H., Jing, N., Li, J.: RemainNet: explore road extraction from remote sensing image using mask image modeling. Remote Sensing. **15**(17), 4215 (2023). https://doi.org/10.3390/rs15174215
8. Luo, L., Wang, J.X., Chen, S.B., Tang, J., Luo, B.: BDTNet: Road extraction by bi-direction transformer from remote sensing images. IEEE Geosci. Remote Sens. Lett. **19**, 1–5 (2022)
9. Yang, Z., Zhou, D., Yang, Y., Zhang, J., Chen, Z.: TransRoadNet: A novel road extraction method for remote sensing images via combining high-level semantic feature and context. IEEE Geosci. Remote Sens. Lett. **19**, 1–5 (2022)
10. Hu, P.C., Chen, S.B., Huang, L.L., Wang, G.Z., Tang, J., Luo, B.: Road extraction by multi-scale deformable transformer from remote sensing images. IEEE Geosci. Remote. Sens. Lett. (2023)
11. Wang, Y., Peng, Y., Li, W., Alexandropoulos, G.C., Yu, J., Ge, D., Xiang, W.: DDU-Net: Dual-decoder-U-Net for road extraction using high resolution remote sensing images. IEEE Trans. Geosci. Remote Sens. **60**, 1–12 (2022)

12. Yan, J., Ji, S., Wei, Y.: A combination of convolutional and graph neural networks for regularized road surface extraction. IEEE Trans. Geosci. Remote Sens. **60**, 1–13 (2022)
13. Chandra, N., Vaidya, H., Ghosh, J.K.: Human cognition based framework for detecting roads from remote sensing images. Geocarto Int. **37**(8), 2365–2384 (2022)
14. Abdollahi, A., Pradhan, B., Alamri, A.: VNet: An end-to-end fully convolutional neural network for road extraction from high resolution remote sensing data. IEEE Access. **8**, 179424–179436 (2020)
15. Wei, Y., Zhang, K., Ji, S.: Simultaneous road surface and centerline extraction from large-scale remote sensing images using CNN-based segmentation and tracing. IEEE Trans. Geosci. Remote Sens. **58**(12), 8919–8931 (2020)
16. Luo, Z., Zhou, K., Tan, Y., Wang, X., Zhu, R., Zhang, L.: AD-RoadNet: an auxiliary-decoding road extraction network improving connectivity while preserving multiscale road details. IEEE J. Selected Top. Appl. Earth Observ. Remote Sens (2023)
17. Yin, A., Ren, C., Yan, Z., Xue, X., Zhou, Y., Liu, Y., Lu, J., Ding, C.: C2S-RoadNet: road extraction model with depth-wise separable convolution and self-attention. Remote Sens. **15**(18), 4531 (2023)
18. Yang, Z.X., You, Z.H., Chen, S.B., Tang, J., Luo, B.: Semi-supervised edge-aware road extraction via cross teaching between CNN and transformer. IEEE J. Selected Top. Appl. Earth Observ. Remote Sens. (2023)
19. 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 Observ. Geoinf. **113**, 102987 (2022). https://doi.org/10.1016/j.jag.2022.102987
20. Christophe, E., Inglada, J.: Robust road extraction for high resolution satellite images. In 2007 IEEE International Conference on Image Processing. IEEE. 5, V-437 (2007, September)
21. https://www.kaggle.com/datasets/balraj98/deepglobe-road-extraction-dataset
22. Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., Hughes, F., Tuia, D., Raskar, R.: Deepglobe 2018: A challenge to parse the earth through satellite images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp .172–181 (2018)
23. https://www.kaggle.com/datasets/hithere016/chn6-roads-dataset
24. Dai, L., Zhang, G., Zhang, R.: RADANet: road augmented deformable attention network for road extraction from complex high-resolution remote-sensing images. IEEE Trans. Geosci. Remote Sens. **61**, 1–13 (2023)
25. Jing, Y., Zhang, T., Liu, Z., Hou, Y., Sun, C.: Swin-ResUNet+: An edge enhancement module for road extraction from remote sensing images. Comput. Vis. Image Understand. 103807 (2023)
26. Li, R., Chen, T., Liu, Y., Jiang, H.: CoupleUNet: Swin Transformer coupling CNNs makes strong contextual encoders for VHR image road extraction. Int. J. Remote Sens. **44**(18), 5788–5813 (2023)
27. Miao, C., Zhang, Z., Tian, Q.: TransLinkNet: LinkNet with transformer for road extraction. In: International Conference on Optics and Machine Vision (ICOMV 2022). SPIE. 12173,138–143 (2022, May)
28. Tao, J., Chen, Z., Sun, Z., Guo, H., Leng, B., Yu, Z., Wang, Y., He, Z., Lei, X., Yang, J.: Seg-Road: a segmentation network for road extraction based on transformer and CNN with connectivity structures. Remote Sens. **15**(6), 1602 (2023)
29. Lan, M., Zhang, Y., Zhang, L., Du, B.: Global context based automatic road segmentation via dilated convolutional neural network. Inf. Sci. **535**, 156–171 (2020)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.