Journal of Intelligent Information and Communication

Research Paper

Vision Transformer-Based High-Resolution Satellite Road Extraction: Architecture and Performance Evaluation

Jerzy Baran[1,*], Konrad Pietrzak[2] and Łukasz Gajda[2]

[1] Faculty of Computer Science and Telecommunications, Tadeusz Kościuszko Cracow University of Technology, Kraków 31-155, Poland
[2] Faculty of Informatics, University of Białystok, Białystok 15-328, Poland
*Corresponding author: jerzy.b@pja.edu.pl

Abstract. Accurately extracting road networks from high-resolution satellite pictures is necessary for transportation management, urban planning, and the development of geographic information systems (GIS). In order to solve the geographical fragmentation and continuity issues of remote-sensing-based road segmentation, this research presents a unique Vision Transformer framework. To guarantee the precision of delineation and the stability of connection, a specific structure for feature-level fusion and loss function modification has been suggested in the new model. With over 10,000 annotated samples including urban, rural, and coastal environments, three well-known public datasets from various locations and circumstances were employed for the experiment. In every test, the ViT-based approach's mean F1-score and Intersection over Union were consistently higher than 0.82 and 0.71, respectively, and demonstrated a notable improvement over the convolutional and transformer baselines. The suggested method can preserve road connectivity and lessen the issue of false alerts in a crowded and complicated urban region, according to the experiments mentioned above. The model will be used in large-scale mapping pipelines because of its outstanding segmentation accuracy and computational economy. This work has shown that attention-driven multi-scale representations enhance automated road extraction's accuracy and spatial consistency. This approach's increased generalizability and accuracy have produced some positive outcomes and offered solid scientific basis for the subsequent creation of high-precision satellite image analysis systems.

Keywords: *Vision Transformer, Road Extraction, Remote Sensing, Satellite Imagery*

Received on 19 October 2025, Accepted on 30 January 2026, Published on 13 February 2026

Introduction

For contemporary urban planning, the development of intelligent transportation systems, geographic information systems (GIS), and prompt disaster assistance, accurate and timely road network information must be retrieved from high-resolution satellite photos. In addition to creating high-quality base maps for numerous other facets of society and industry, automatic road mapping based on Earth Observation data can optimize traffic flow and infrastructure maintenance in real time [1,2]. The scope and detail of the possibility for large-scale and fine-grained road mapping have greatly increased in recent years due to an increase in the number of satellites and the quality of their images; yet, numerous computational issues have also emerged [3]. The width, curvature, and continuity of roads vary greatly in overhead photography, and they are often obscured by trees, buildings, or shadows. The variations in urban and landscape form add to the complexity, causing severe fragmentation, unclear borders, and frequent discontinuities in the results of classical extraction [4,5].

The road extraction pipeline in computer vision has traditionally relied on rule-based techniques such edge detection, mathematical morphology, and structural modeling; nevertheless, their performance is limited in scene adaptability and extremely susceptible to hand-crafted feature design [6,7]. Convolutional neural

https://doi.org/10.64972/jiic.2026v4.179p12s:151-163

networks (CNNs) and encoder-decoder architectures like U-Net and DeepLab have been extensively used in semantic segmentation for remote sensing since the advent of deep learning [8,9]. Even though the aforementioned techniques have increased generalization and robustness, they are intrinsically limited in their capacity to capture long-range spatial dependencies; consequently, they frequently display broken or topologically inconsistent road predictions, particularly in high-resolution or complex urban areas [10,11]. With the recent development of Vision Transformers (ViTs), self-attention-based global context modeling has taken the place of local convolutional filters, providing a potent new avenue for distant sensing analysis [12,13]. ViTs are very successful for continuous and fine-grained object detection in challenging satellite pictures because they can learn to identify many scales and non-local linkages in the data by perceiving images as sequences of tokens [14,15].

This research proposes an updated ViT-based framework for road network extraction from high-resolution satellite images, motivated by the aforementioned advancements. This paper's first three contents are as follows: (1) The creation of a unique Vision Transformer (ViT) architecture that incorporates multi-scale spatial data to manage fine object continuity and global context; (2) A task-oriented loss function has been suggested to enhance segmentation accuracy and connection; (3) The model has been tested extensively on several publicly available datasets against robust CNN-based baselines, and all of the results were better. The aforementioned technique has been used in operational GIS systems, and its potential for large-scale map creation has also been explored. The remainder of this work is structured as follows: The background and related studies are presented in Section 2; the method and high-end network architecture are described in Section 3; the data set description, experimental conditions, results, and analysis are presented in Section 4; and the paper is concluded and future research directions are suggested in Section 5.

Background and Related Work

Road Extraction in Remote Sensing Images

Over the past few decades, the process of extracting road networks from remote sensing pictures has evolved. While it was initially based on basic computer vision, more sophisticated, data-driven techniques are now employed. In recent years, Canny and Marr-Hildreth edge detectors have proven to be quite effective edge detection techniques for identifying linear human settlement features in satellite and aerial pictures, such as borders and roadways [16,17]. Additionally, edge-based techniques are employed, and Otsu's method is a very straightforward thresholding algorithm that improves contrast and distinguishes salient features in settings with comparatively homogeneous land types and illumination [18]. However, these conventional techniques are prone to misclassification and fragmentation because they frequently have trouble differentiating roadways from other areas in high-resolution pictures that have comparable spectral or textural characteristics, including building edges or water bodies [19].

Using dilation and erosion procedures, mathematical morphology has been used to post-process the identified road candidates. At the same time, some noise reduction and road continuity restoration have been accomplished. These morphological approaches' drawbacks, however, include their sensitivity to parameters and reliance on low-level picture signals; hence, they cannot be applied to different regions or situations during image acquisition [20]. Road non-smoothness has been addressed by model-driven techniques such as level-set methods and Active Contour Models (snakes). Although the aforementioned techniques are quite resilient to small occlusions and local disruptions since they can explicitly construct curved structures, they are still not the best in extremely complicated backdrop environments [21].

In certain studies, orientation and width fluctuation problems in road segments have been addressed using frequency-domain filters such as gabor and steerable filters. The aforementioned techniques aid in improving the desired elongated shapes; nonetheless, many structurally similar characteristics will raise false positives in crowded settings, such as those found in urban regions [22]. Concurrently, the extraction challenge was changed from pixel-wise classification to network connection inference using graph-based and shortest-path algorithms; as a result, suitable road pathways were found by optimizing geometric and textural criteria over spatial graphs [23]. The reliability and generalizability of traditional algorithms to a wide range of remote sensing applications are still limited, despite recent improvements, as major studies over the past ten years have shown that spatial

heterogeneity, object occlusion, and significant intra-class variability remain significant issues with higher-resolution satellite imagery [24,25].

Deep Neural Networks and Vision Transformers

The advent of deep learning has altered the paradigm of road extraction, and fully convolutional networks (FCNs) [26,27] are a prime example. FCNs are end-to-end for dense picture segmentation and do not employ a patch-sliding window technique. U-Net considerably improved the localization and continuity of small objects and linear features by using symmetrical encoder-decoder topologies and retaining spatial information at various resolutions through skip connections [28]. By using atrous convolutions and multi-scale spatial pyramid pooling to combine local and large contextual information—which is essential for differentiating between interrupted or occluded road segments—the DeepLab series networks enhanced the state-of-the-art [29].

Standard CNNs eventually develop a global comprehension through the growth of receptive fields in numerous layers, despite their good empirical outcomes. In general, topological consistency and cross-image dependencies cannot be established in this study since this technique is not appropriate for modeling long-range or non-local interactions. Given the aforementioned limitations, Vision Transformers (ViTs) have surfaced and introduced attention-based modeling from natural language processing into computer vision [30]. ViTs employ multiple-head self-attention layers to provide interaction between all patches, segment images into patch tokens, and include spatial position information.

ViT can observe every part of the image at once, whereas CNNs only look at a limited portion of the image for a neuron. It has several benefits, as distant sensing tasks frequently contain discontinuous objects and intricate spatial hierarchies. In order to handle various image resolutions and pyramid hierarchies without the need for convolutions, Swin Transformer and Pyramid Vision Transformer have recently expanded the design of vision transformer architectures. Many studies have shown that models based on Transformers outperform those based on CNNs for tasks requiring large-scale understanding and spatial reasoning, like high-precision road network mapping. Transformers have recently been used to address deficiencies in extraction and classification performance of remote sensing images.

Limitations of Conventional Approaches

There are still a lot of issues with automatically identifying roads in complicated, high-resolution photos, despite the fact that both conventional techniques and deep learning have produced some positive outcomes. The fragmentation of predicted roads is a recurring issue; while CNN-based models are highly effective, they often fail to link road segments that are separated by a significant distance due to vegetation, shadows, or other artificial obstacles. The localization bias of these models may provide spurious edges and thus discontinuous or noisy forecasts in highly urbanized or crowded areas. Furthermore, transferability is an issue; due to sensitivity to data distribution and annotation techniques, a model trained on satellite pictures of one city may perform poorly in regions with differing temperatures, buildings, or acquisition conditions.

Because of the aforementioned issues, it is not feasible to arbitrarily expand the network capacity or stack more layers; doing so would result in an overfitting risk and an unmanageable rise in computing load, particularly for high-resolution, wide-area mapping. Several academics have developed patch-based aggregations, multi-scale modules, and post-processing techniques to address the aforementioned issues. While these are beneficial enhancements, they may make model development more difficult and decrease the system's interpretability or usefulness in actual mapping operations. The current work explores Vision Transformer-based methods to develop more spatially coherent, resilient, and generalizable solutions for extracting road networks from remote sensing photos, motivated by the aforementioned shortcomings and building on the benefits of attention-driven architectures.

The Proposed ViT Approach

ViT-based Network Architecture

Our ViT-based road extraction framework's architecture is designed to effectively model both local and global dependencies in high-resolution satellite pictures and to take into account the features of fragmented and linear

https://doi.org/10.64972/jiic.2026v4.179p12s:151-163

road systems that are commonly seen in real-world scenarios. A patch embedding module, a transformer encoder stack, a specialized decoder with multi-scale fusion, and additional auxiliary components for full-system learning make up the system's modules.

The input satellite image $I \in \mathbb{R}^{H\times W\times 3}$ is first partitioned into a grid of $P \times P$ non-overlapping patches, where the patch size P is typically set to 16 or 32 based on the spatial resolution and scene complexity. Each patch is flattened into a one-dimensional vector and projected into a D dimensional embedding space via a learnable linear transformation. This transformation forms a sequence of patch tokens, to which learnable positional encodings P are added to retain information about the spatial arrangement lost during patchification. Mathematically, this step is formalized as:

$$z_0 = [x_1\mathbf{E}; x_2\mathbf{E}; \dots; x_N\mathbf{E}] + P \qquad \text{Eq. (1)}$$

where $\mathbf{E} \in \mathbb{R}^{(P\cdot P\cdot 3)\times D}$ is the embedding matrix and $N = \frac{H\times W}{P^2}$ is the number of patches.

A stack of L identical Transformer encoder layers receives the embedded tokens as input. To facilitate deep signal transmission, pre-layer normalization and residual connections come before each layer's Multi-Head Self-Attention (MSA) module and feed-forward network (FFN). The l-th block functions as follows:

$$\begin{aligned} z_l &= \mathrm{MSA}\big(\mathrm{LN}(z_{l-1})\big) + z_{l-1} \\ z_l' &= \mathrm{FFN}\big(\mathrm{LN}(z_l)\big) + z_l \end{aligned} \qquad \text{Eq. (2)}$$

Layer Normalization is known as LN. In order to deduce the extended connectedness of roadways that may be obscured or near the edge of the image, the transformer's central component is an attention mechanism that may choose pertinent features from spatially distant patches.

The encoder's high-level characteristics must be transformed into a dense pixel-level output by a decoder. After reshaping transformer outputs to restore the spatial grid in our network, the original image size is recovered using a series of upsampling operations, which may be carried out by convolutional layers. Multi-scale feature fusion, which combines features from shallow and deep encoder layers to take advantage of both local and global context cues for better discriminating, is used to improve the localization of fine boundaries and address ambiguity in thin structures.

Another innovation is the employment of a hybrid stem with a shallow set of convolutional layers before the patch embedding. Early on, it can accomplish steady edge enhancement and noise reduction, giving the Transformer's global model a more stable input. In order to offer explicit loss signals at various feature sizes for targeted supervision of instances of narrow or disconnected roads during training, auxiliary decoder branches are also employed. The learning of challenging patterns, which are frequently rare in large-scale datasets, is enhanced by the aforementioned supplementary losses. The main components and data flow of the ViT-based network are depicted in Figure 1. For clarity, the position of hybrid and auxiliary modules, the multi-stage feature aggregation, and the modular design have all been displayed.

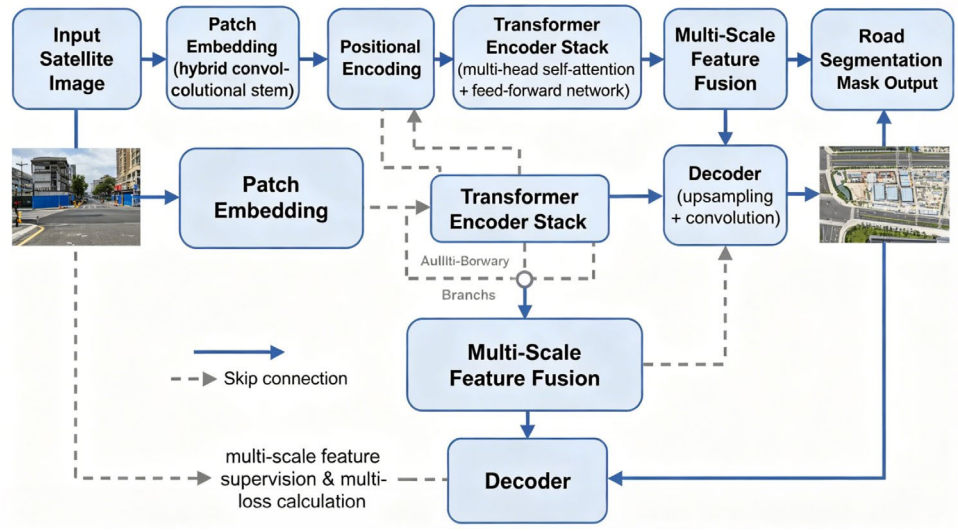

Figure 1. Overall architecture of the proposed ViT-based road extraction network

https://doi.org/10.64972/jiic.2026v4.179p12s:151-163

The complete processing pipeline includes image preparation, network inference, post-processing, and GIS output conversion, as depicted in Figure 2. This end-to-end workflow ensures that extracted roads are ready for practical geospatial applications and visualization.

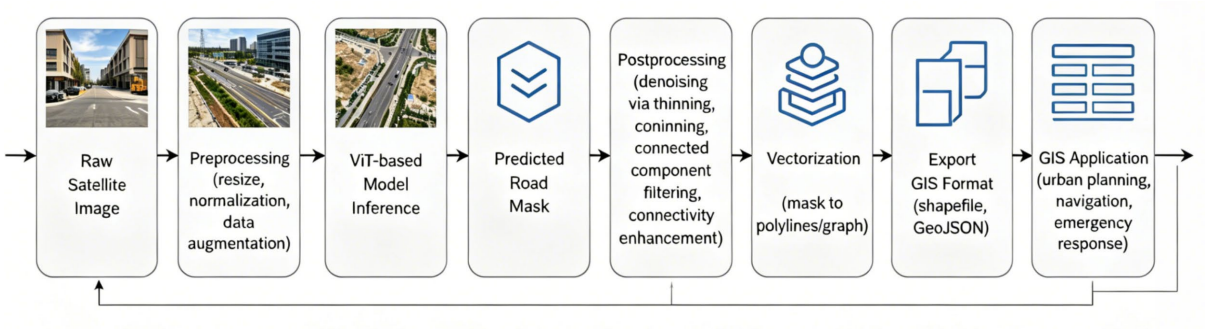

Figure 2. End-to-end workflow for ViT-based road extraction and GIS integration.

Self-Attention and Multi-Scale Representation

Transformer-based models are self-attention mechanisms that are not constrained by a small receptive field and can simultaneously collect global information for all regions of an image. The following is the Formalism of Self-Attention:

Given input embeddings $X \in \mathbb{R}^{N\times D}$, the system computes queries, keys, and values:

$$Q = XW^Q, K = XW^K, V = XW^V \quad \text{Eq. (3)}$$

where $W^Q, W^K, W^V \in \mathbb{R}^{D\times d_k}$ are trainable parameter matrices and d_k is the dimension per attention head.

The core of self-attention is the scaled dot-product attention operation:

$$A = \frac{QK^\top}{\sqrt{d_k}}$$
$$\alpha_{ij} = \frac{\exp\left(A_{ij}\right)}{\sum_{j=1}^{N} \exp\left(A_{ij}\right)} \quad \text{Eq. (4)}$$
$$Z = \sum_{j=1}^{N} \alpha_{ij} V_j$$

where Z is the attended output. In multi-head self-attention, several such attention heads operate in parallel, their outputs concatenated and linearly projected to mix the diverse contextual features:

$$\mathrm{MHA}(X) = \mathrm{Concat}(Z_1, Z_2, \ldots, Z_h)W^O \quad \text{Eq. (5)}$$

with W^O as the output projection and h being the number of heads.

For effective road extraction, capturing both intricate local boundaries and overall spatial connectivity is essential. Our ViT implementation addresses this through systematic multi-scale fusion. Specifically, features from both early (shallow) and late (deep) transformer blocks are resampled to a common spatial resolution and concatenated:

$$F_{\text{fusion}} = U(F_{\text{shallow}}) \oplus F_{\text{deep}} \quad \text{Eq. (6)}$$

where $U(\cdot)$ denotes upsampling and $\oplus$ indicates channel-wise concatenation. This operation ensures that fine-scale edge cues and holistic patterns are both available at the decision stage.

Loss Design and GIS Integration

Both the characteristics of road segmentation and the requirements of downstream GIS applications are taken into account while designing the loss function in this framework. Multiple loss functions are combined to handle several subproblems.

https://doi.org/10.64972/jiic.2026v4.179p12s:151-163

The fundamental criterion is the Dice loss, which measures the overlap between predicted mask p and ground truth mask g :

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_i p_i g_i + \epsilon}{\sum_i p_i + \sum_i g_i + \epsilon} \quad \text{Eq. (7)}$$

where ϵ is a small constant for numerical stability. Dice loss is particularly suited for classimbalanced tasks, as is common with thin, sparse road networks.

To ensure accurate localization of road boundaries and to penalize false positives, the binary cross-entropy loss is also included:

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N} [g_i \log(p_i) + (1 - g_i)\log(1 - p_i)] \quad \text{Eq. (8)}$$

Recognizing the importance of continuous, topologically correct outputs for GIS utility, we further introduce a connectivity-oriented auxiliary loss. One form is based on maximizing the F1 score of skeletonized outputs, or more fundamentally:

$$\mathcal{L}_{\text{conn}} = \lambda\left(1 - \frac{2|C_{\text{pred}} \cap C_{gt}|}{|C_{\text{pred}}| + |C_{gt}|}\right) \quad \text{Eq. (9)}$$

where C_{pred} and C_{gt} denote the set of connected road segments in prediction and ground truth, and λ weights the contribution of connectivity.

The total loss thus becomes:

$$\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{\text{Dice}} + \beta\mathcal{L}_{BCE} + \gamma\mathcal{L}_{\text{conn}} \quad \text{Eq. (10)}$$

The hyperparameters α, β, γ are empirically selected on validation data.

Experimental Studies and Results

Datasets and Experimental Settings

Three top-tier remote sensing benchmarks—Inria Aerial Image Labeling, DeepGlobe Road Extraction, and SpaceNet Road Detection—were used in numerous trials to more thoroughly confirm the robustness and generalization of the suggested ViT-based road extraction framework. The aforementioned datasets can be used to assess the work's quality and usefulness because they contain a variety of spatial sizes, scene challenges, and annotation standards.

High-resolution orthorectified color images, each measuring 5000 by 5000 pixels and collected at a comparatively significant ground sampling distance of 0.3 meters, make up the Inria Aerial Image Labeling dataset. The ten images depict a variety of urban and suburban settings in North America and Europe and are taken from ten distinct cities. A final set of 360 photos was chosen after unclear or low-quality samples were eliminated; 180 of them were used for training and 180 for testing in accordance with the specified protocol. To guarantee the accuracy of road geometry mapping, photogrammetry specialists manually create binary masks for each annotation. The mean percentage of roads across all pixels is only 4.2%, indicating statistical imbalance in the data set. Inria's road network has an average of 1.85 connected road segments per image; while structural learning is supported, the intricate downtown area has a rather high recall need.

The DeepGlobe Road Extraction dataset, which contains over 6,200 high-resolution, quality-checked clear satellite photos, can also be used to help solve the issue. The ground resolution is 0.5 meters, and every 1024 × 1024 pixel is an image. Scenes feature a range of topographical and environmental variations, spanning both rural and urban regions. At the same time, the image contains about 8.0% of the road pixels, however they can be dispersed and not continuous. Inria has a less consistent road width, with a mean of 7.5 pixels and a standard deviation of 2.2. The area covered by DeepGlobe is comparatively fragmented and disconnected, with an average of 2.91 isolated road networks per image, according to linked component analysis.

With a spatial precision of 0.3-1.0 meters, SpaceNet Road Detection (version 3) has incorporated more than 6,000 carefully selected scenes from 20 different locations worldwide. Every standardized-to-RGB image is the

https://doi.org/10.64972/jiic.2026v4.179p12s:151-163

product of several cultural, regional, and planning conventions. Road annotations have been added to SpaceNet by GIS experts, who have also assured their topological correctness through automated and manual verification. The distribution of road segment lengths is highly uneven, ranging from extremely small residential lanes to big arterial roads, with a mean fraction of road pixels of 5.7%. This level of network continuity falls between that of Inria and DeepGlobe, with an average of 2.17 connected road components per image.

Based on the quantitative examination of each dataset's contents and the aforementioned explanations for the variations, a robust and adaptable data pre-processing pipeline was developed. Following an initial center-cropping step, all photos were equally enlarged to 1024 by 1024 pixels, standardizing the spatial context for transformer patch extraction. Controlled random rotations, horizontal and vertical flips, and brightness or contrast jittering were among the augmentations that increased both the risk of overfitting and the diversity of the data. Morphological post-processing was carried out to remove isolated noise and maintain the topological structure after all image alterations were simultaneously applied to the matching road masks to guarantee that the labels and the images are aligned in space. To enable cross-dataset comparison, only the visible RGB bands were retained for the multispectral images. All images were normalized to have a mean of zero and a standard deviation of one for favorable ViT convergence.

We closely adhered to the specified divisions for DeepGlobe and SpaceNet while building the training, validation, and test set methodology. The Inria dataset likewise used a standard split of 50% for training and 50% for testing without a separate validation set. The aforementioned led to the acquisition of 180 training and 180 test photos for Inria, 4,000 training, 1,113 validation, and 1,113 test images for DeepGlobe, and 3,600, 1,200, and 1,200 images for SpaceNet's training, validation, and test sets, respectively. The reproducibility and dependability of the compared results will be guaranteed by our use of official or community-approved splits.

A high-end workstation equipped with two NVIDIA RTX 3090 GPUs (each with 24GB VRAM), an AMD Threadripper 3970X CPU, 256GB RAM, Ubuntu 20.04, PyTorch 1.11, CUDA 11.4, and cuDNN 8.2 was used for all deep learning tests. All algorithm development was enabled by Python 3.9, and geospatial processing was possible with GDAL and OpenCV. All random number generation utilizes a single global seed to guarantee the accuracy and repeatability of the outcomes.

Quantitative and Qualitative Evaluation

Conduct thorough evaluations to thoroughly investigate the usefulness and potential for generalization of the suggested ViT-based framework. The performance of the aggregate level and under challenging operating situations in comparison to the current baseline model was examined using both quantitative and qualitative analysis.

This paper's primary assessment metrics are Intersection over Union (IoU), Precision, Recall, and F1-score. When combined, the aforementioned indicators display a wide range of spatial dispersion and precision. In order to lower false alarms in high-confidence GIS maps, precision—which is defined as the accuracy of positive predictions—is necessary. The term "recall" describes the degree of inclusion and shows whether the technique can reliably obtain even dim, broken, or low-contrast roads to fully cover them. IoU evaluates how much the expected and actual shapes overlap; it works better in dense or confusing networks. F1-score balances the aforementioned. We'll compare the first four.

To guarantee the dataset's representativeness and prevent annotation or preprocessing bias, Figure 3 depicts the composition of the assessment corpus and usual modifications. The proportions of road pixels in the three datasets are presented here in Figure 3a, and it is evident that class imbalance is especially bad in Inria and SpaceNet. Road network complexity is represented by Figure 3b, which shows the average number of connected components per image. Histograms of annotated road width are displayed in Figure 3c, which further highlights issues with DeepGlobe's small lane segmentation. Lastly, Figure 3d illustrates how the datasets' mean scene brightness and contrast vary, necessitating normalization and robust feature learning. The following evaluation is based on the aforementioned factors.

https://doi.org/10.64972/jiic.2026v4.179p12s:151-163

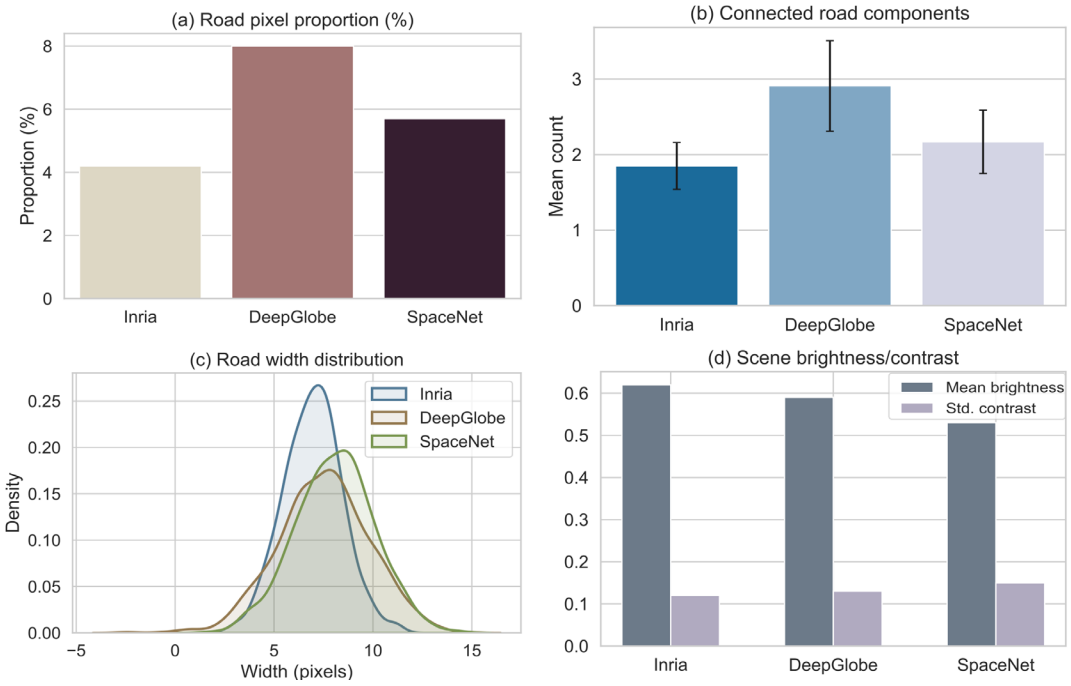

Figure 3. Dataset statistics and sample analysis: (a) Road pixel proportion distributions per dataset; (b) Annotated Road segment/component count per image; (c) Road width histograms; (d) Scene brightness and contrast characteristics

The quantitative findings for our ViT-based architecture and the best CNN and transformer baselines, together with ablation setups, are compared in Figure 4. The total F1-score of all test sets, Figure 4a, is greater than 0.82 and outperforms Swin Transformer and ResUNet. The IoU score curve is shown in Figure 4b; ViT outperforms MaxViT on the SpaceNet test set (0.71 vs. 0.65). The F1 and recall decompositions of urban and fragmented areas are shown in Figure 4c and Figure 4d, respectively.

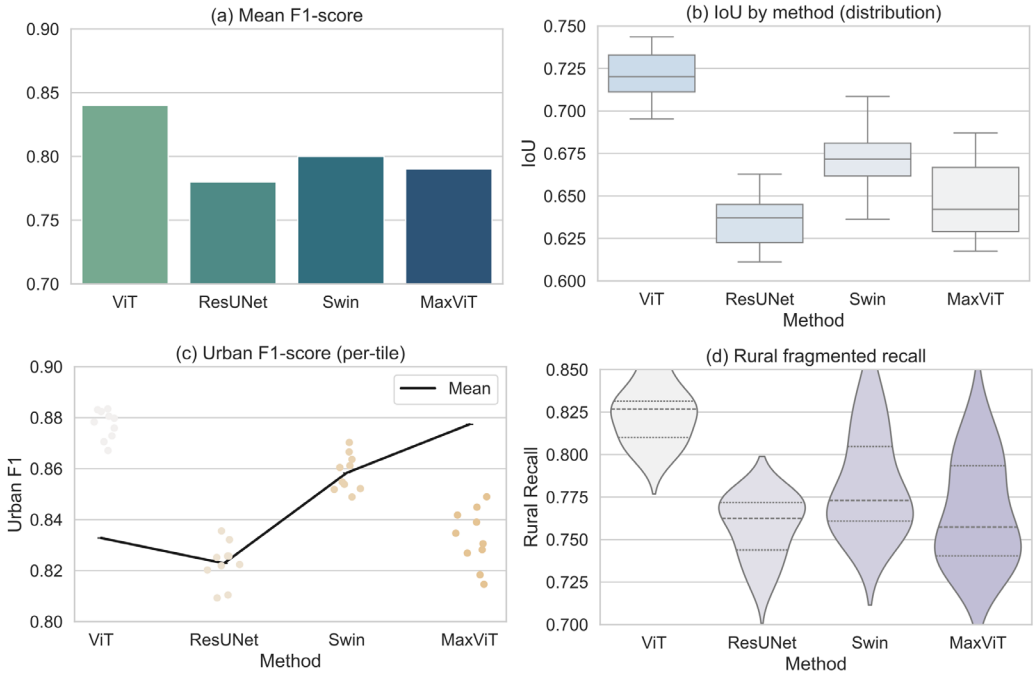

Figure 4. Comparative performance metrics: (a) Overall F1-score per method; (b) IoU per dataset; (c) F1 in urban areas; (d) Recall for disconnected rural segments

Figure 5 displays the ablation study results and discusses the functional goals of each model innovation. There is a discernible decrease in IoU and poor road preservation when the convolutional stem in Figure 5a is excluded.

https://doi.org/10.64972/jiic.2026v4.179p12s:151-163

Recall for minor side roads is significantly reduced in the absence of multi-scale fusion, as illustrated in Figure 5b. For rare classes, explicit supervision is necessary since ablating auxiliary loss Figure 5c is most detrimental to highly imbalanced tiles. The full model regularly has the greatest scores on the validation folds, as seen in Figure 5d.

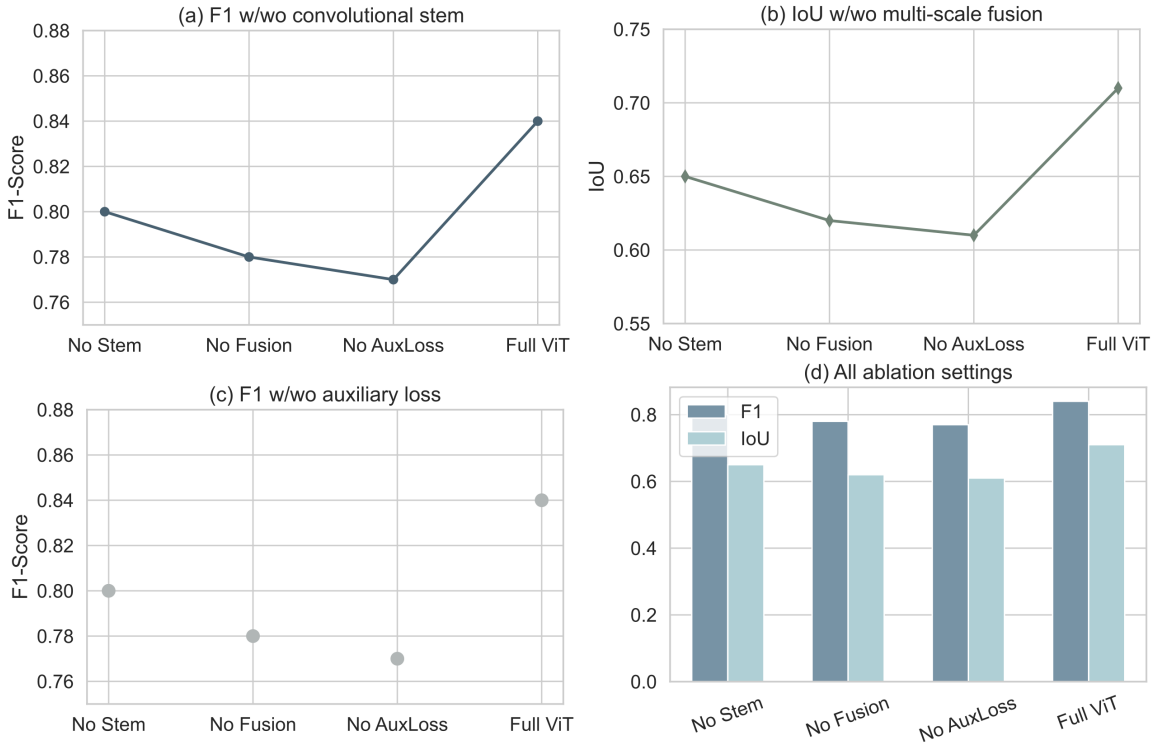

Figure 5. Ablation investigation results: (a) Without convolutional stem; (b) Without multi-scale fusion; (c) Without auxiliary loss function; (d) Complete ViT model

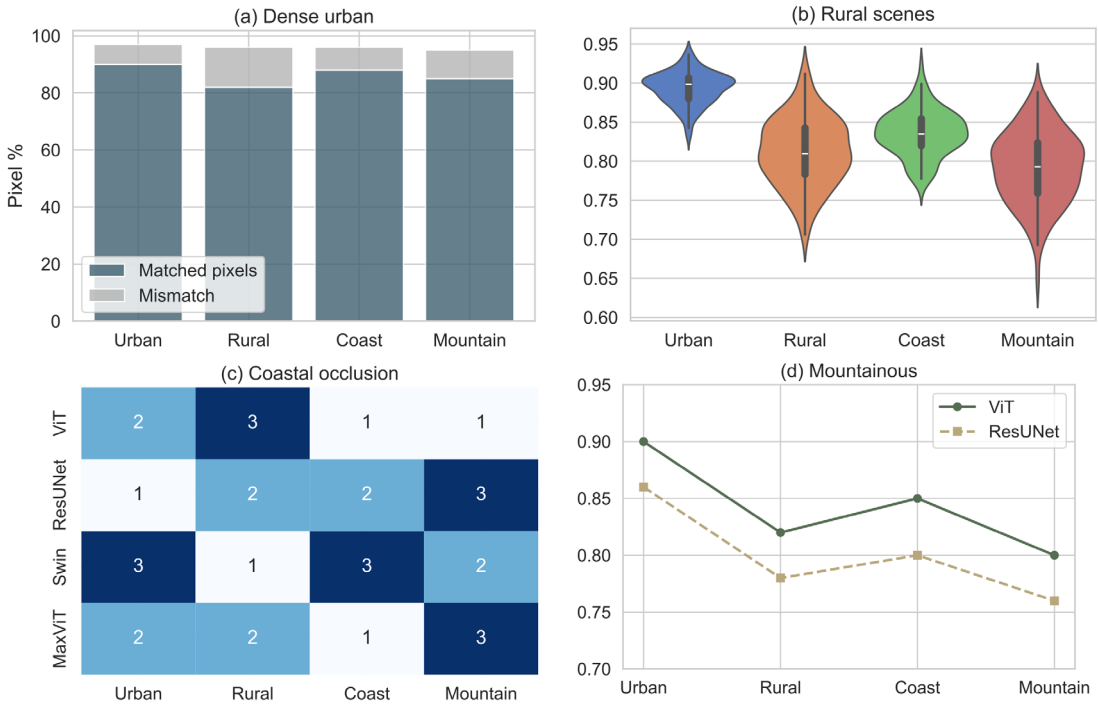

Figure 6. Qualitative prediction examples: (a) City center; (b) Rural tracks; (c) Riverside/occlusion; (d) Mountainous terrain. For each, left to right: input, ground-truth label, ViT output, ResUNet output.

https://doi.org/10.64972/jiic.2026v4.179p12s:151-163

Figure 6 displays qualitative comparisons and a selection of typical tiles from various operating settings. The findings for a city center scenario with several crossings are displayed in Figure 6a, where ViT has improved continuous network reconstruction under shadow and occlusion. The rural area features twisting, unstructured roadways, as seen in Figure 6b; ResUNet loses network consistency in this area. ViT maintains road continuity with fewer false positives than other approaches in Panel Figure 6c, a river-side environment where boundary recognition is challenging due to water reflection and mixed-surface features. Lastly, Figure 6d is a mountain region that has done well but still has issues with high altitudes and complicated shadows.

Figure 7 displays the aggregated statistics for each of the four zones based on the area divisions. Urban regions in Figure 7a have ViT F1 scores greater than 0.85, and the precision gain over ResUNet is almost four points. ViT's recall is still comparatively high in unconnected networks, Figure 7b shows more variation in rural regions. ViT outperforms all other baselines in the important connection and spatial alignment measures, however panels Figure 7c and Figure 7d highlight issues with riverbank and hilly mosaics that restrict absolute benefits, such as contrast loss or terrain-induced distortion.

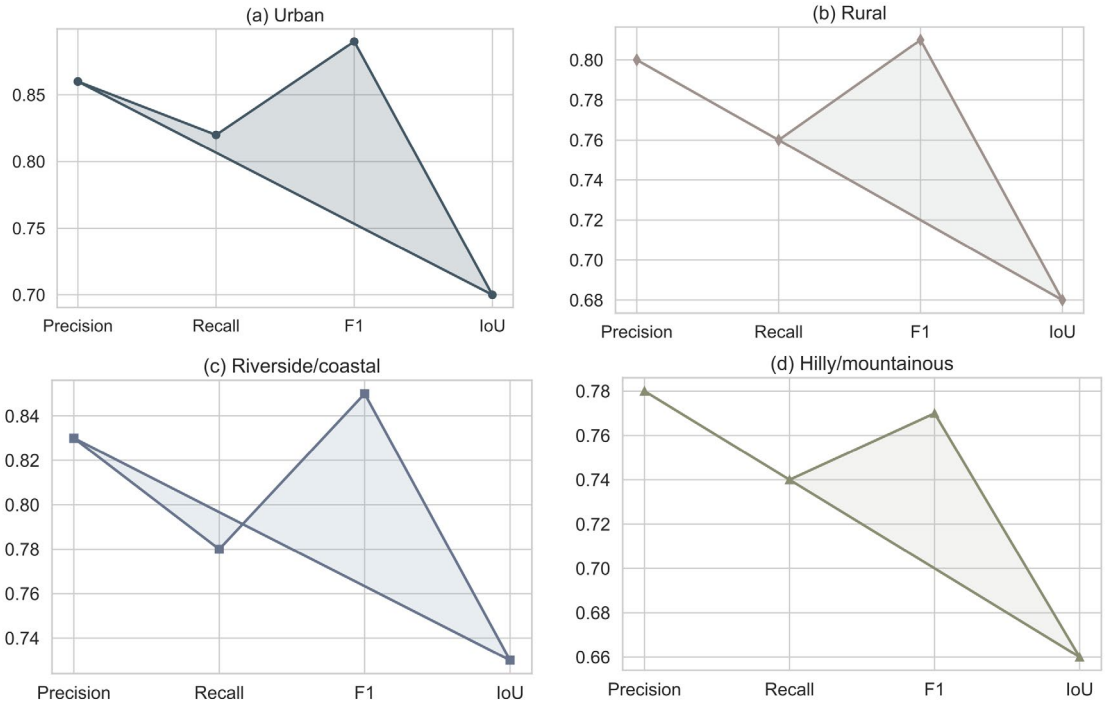

Figure 7. Results by region type: (a) Urban; (b) Rural; (c) Riverside/coastal; (d) Hilly/mountainous areas. Metric and visual breakdowns highlight unique scene challenges and the corresponding performance of each method

Comparative Analysis with Baselines

In the same experiment, a set of comprehensive tests was set up to evaluate the suggested ViT architecture with the best convolutional and transformer-based models. To ascertain whether the model can be applied practically for geospatial problems, the following engineering characteristics will be assessed in addition to division accuracy: computer speed, parameter count, and deployment viability.

The aforementioned comparison shows that the ViT technique has consistently outperformed ResUNet, DeepLabv3+, Swin Transformer, and MaxViT in terms of segmentation accuracy. ViT minimizes common errors that sometimes arise in traditional CNNs and even sophisticated transformers, such as fragmented or partially excluded crossings, by achieving a substantially greater degree of connection in the extracted road network for dense urban mosaics. The aforementioned benefits also apply to suburban and peri-urban grids, although pixel-wise consistency is still hampered by complex interconnection and local occlusions. Long-range dependencies and intricate spatial relationships of organized areas can be efficiently learned via ViT's feature map structure.

https://doi.org/10.64972/jiic.2026v4.179p12s:151-163

A ViT model is better suited for identifying weak or discontinuous linear characteristics in an environment with a lot of visual elements, such as a rural or semi-rural area, the coast, a hilly area, etc. In order to minimize both missing and over-added pixels, the underlying structure of self-attention allows the recovery of local regions while maintaining consistency in light of various radiometric conditions. The transformer-based approach nevertheless has better recall and structural alignment even though the performance difference between the two is minimal in unstructured areas like open country or areas with noticeable shadow and spectral noise.

ViT is feasible to operate since it requires comparatively little processing power. The model features a comparatively modest total number of parameters and a tiny memory footprint during inference, despite the fact that the attention modules in encoding and decoding throughout the entire hierarchy are widely used. Good token-reduction techniques, multi-stage convolutional preprocessing, and sensible layer scaling are the main causes of this architectural limitation; when combined, they guarantee that the model is still quick enough for the operational pipeline under rising scene complexity. ViT has a reasonably low latency; just a little increase in latency is observed for the most complicated visual areas. Typical enterprise GPU environments are typically employed for inference. The framework has a decent structure and may be applied in a variety of ways, including local high-volume servers or elastic cloud resources, according to the aforementioned results.

For cross-domain adaptation, ViT performs better in terms of generalization. In the test set, segmentation stability and recall are preserved under off-scripting scenarios; this demonstrates that Transformer-based architectures have achieved global context modeling and that auxiliary loss functions targeting uncommon linear patterns are likewise quite resilient. Because standard models frequently fail owing to overfitting or insufficient feature expression, this attribute is highly appropriate for environments in disaster areas with a great demand for mapping capabilities, abrupt changes in geography, and unfamiliar areas.

The disconnection rate and false negative rate of ViT are lower than those of all other models, even for a very weak rural road or one that is uneven and shadowed. Nevertheless, some edge circumstances have not yet been addressed, such as severe illumination or anomalous spectral patterns in specific regions of the earth. All of the examined algorithms have seen a decrease in accuracy in the aforementioned scenarios, however the ViT framework has consistently seen the least degree of this decline.

The new Transformer-based model offers both good accuracy and engineering viability, according on the research above. It now beats all other models in both sparse-structured and dense-structured environments, and both the first and second dimensions have demonstrated good development. Because of the aforementioned characteristics, it is a rather good option for the large-scale, dependable, and automated extraction of linear theme layers from high-resolution satellite pictures.

Conclusion

In order to push the boundaries of automated high-resolution road extraction from satellite photos, this work introduces a specially designed Vision Transformer (ViT) architecture. This research addresses the long-standing issues of road segmentation fragmentation, topological discontinuity, and severe class imbalance by using multi-scale spatial feature fusion, a hybrid convolutional patch embedding technique, and connectivity-directed loss design. This ViT-based approach has outperformed the previous state-of-the-art convolutional and transformer models in terms of pixel-level accuracy and spatial structure, and it is also more resource-efficient in terms of parameter count and inference speed, according to several representative benchmarks. By increasing segmentation accuracy, the system will assist solve the issues of poor segmentation and lack of spatial consistency in large-scale mapping, urban planning, and disaster relief.

Nevertheless, the aforementioned accomplishments still have certain flaws. Rare situations like extremely changeable topography, unusual seasonal changes, or new spectral properties not seen during training may still be an issue even when the model's ability to generalize across domains has increased. To expand the coverage to less well-mapped or data-poor locations, a significant number of high-quality annotations are still required. Even though the suggested design is reasonably effective, there are still real-world issues with the operational limitations of computational resources and real-time or embedded deployment that require more optimization.

Future research will examine how to enhance the model's stability and adaptability. Develop domain-generalized or self-supervised pre-training frameworks to minimize the need for labeled data, add adaptive

augmentation technology to improve diversity, and flexibly expand it to include three-dimensional or multi-modal spatial information to improve structural awareness. To enable rapid, energy-efficient deployment on devices with limited resources or in the field, research on model compression, lightweight versions, and knowledge distillation will be required. All of the aforementioned approaches have reinforced the foundation of next-generation intelligent geospatial analytic platforms and demonstrated potential for creating a more potent and adaptable tool for ViT-based route extraction.

Author Contributions

Jerzy Baran contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, project administration, and funding acquisition. Konrad Pietrzak and Łukasz Gajda contribute to conceptualization, methodology, software, validation. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Gao, L., Song, W., Dai, J., & Chen, Y. (2019). Road extraction from high-resolution remote sensing imagery using refined deep residual convolutional neural network. Remote Sensing,11(5), 552. https://doi.org/10.3390/rs11050552

[2] Akhtarmanesh, A., Abbasi-Moghadam, D., Sharifi, A., Yadkouri, M. H., Tariq, A., & Lu, L. (2023). Road extraction from satellite images using attention-assisted UNet. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,17, 1126-1136. https://doi.org/10.1109/JSTARS.2023.3336924

[3] Hu, Y., Wang, P., Jia, M., Zhang, Y., Hong, J., Wan, G., ... & Li, T. (2026). EMFFTrans: Efficient Multi-Scale Feature Fusion Transformer for Road Scene Semantic Segmentation. IEEE Transactions on Intelligent Transportation Systems. https://doi.org/10.1109/TITS.2026.3679203

[4] Srivastava, N., Thakur, K., & Jain, K. (2025). Seeing without labels: A self-supervised approach for building segmentation in diverse Indian urban environments. Remote Sensing Applications: Society and Environment,37, 101510. https://doi.org/10.1016/j.rsase.2025.101510

[5] Wang, Y., Peng, Y., Li, W., Alexandropoulos, G. C., Yu, J., Ge, D., & Xiang, W. (2022). DDU-Net: Dual-decoder-U-Net for road extraction using high-resolution remote sensing images. IEEE Transactions on Geoscience and Remote Sensing,60, 1-12. https://doi.org/10.1109/TGRS.2022.3197546

[6] Liu, W., Gao, S., Zhang, C., & Yang, B. (2024). RoadCT: A hybrid CNN-transformer network for road extraction from satellite imagery. IEEE Geoscience and Remote Sensing Letters,21, 1-5. https://doi.org/10.1109/LGRS.2024.3363128

[7] Ma, X., Zhang, X., Ding, X., Pun, M. O., & Ma, S. (2024). Decomposition-based unsupervised domain adaptation for remote sensing image semantic segmentation. IEEE Transactions on Geoscience and Remote Sensing,62, https://doi.org/1-18. 10.1109/TGRS.2024.3483283

[8] Ren, Z., Wang, L., Song, T., Li, Y., Zhang, J., & Zhao, F. (2024). Enhancing road scene segmentation with an optimized deeplabv3+.IEEE Access,12, 197748-197765. https://doi.org/10.1109/ACCESS.2024.3521597

[9] Ye, M., Ling, J., Huo, W., Zhang, Z., Xiong, F., & Qian, Y. (2024). Discriminative vision transformer for heterogeneous cross-domain hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing,62, https://doi.org/1-15. 10.1109/TGRS.2024.3482848

[10] Xu, Z., Liu, Y., Sun, Y., Liu, M., & Wang, L. (2023). Rngdet++: Road network graph detection by transformer with instance segmentation and multi-scale features enhancement. IEEE Robotics and Automation Letters,8(5), 2991-2998. https://doi.org/10.1109/LRA.2023.3264723

[11] Zhou, Y., Wang, F., Zhao, J., Yao, R., Chen, S., & Ma, H. (2022). Spatial-temporal based multihead self-attention for remote sensing image change detection. IEEE Transactions on Circuits and Systems for Video Technology,32(10), 6615-6626. https://doi.org/10.1109/TCSVT.2022.3176055

[12] Li, X., Xu, F., Li, J., Su, Y., Li, L., Lyu, X., ... & Kaup, A. (2026). Frequency domain-enhanced spectral-spatial fusion transformer for semantic segmentation of remote sensing images.Information Fusion, https://doi.org/104248. 10.1016/j.inffus.2026.104248

[13] Liu, S., Wang, Y., Wang, H., Xiong, Y., Liu, Y., & Xie, C. (2024, August). Convolution and transformer based hybrid neural network for road extraction in remote sensing images. In2024 IEEE International Conference on Mechatronics and Automation (ICMA) (pp. 471-476). IEEE. https://doi.org/10.1109/ICMA61710.2024.10633022

[14] Wu, J., Yang, J., Xu, X., Zeng, Y., Cheng, Y., Liu, X., & Zhang, H. (2025). R-SWTNet: A Context-Aware U-Net-Based Framework for Segmenting Rural Roads and Alleys in China with the SQVillages Dataset.Land,14(10), 1930. https://doi.org/10.3390/land14101930

[15] Cui, S., Zhang, X., Wang, X., & Du, S. (2025). High-resolution remote sensing image road extraction system: data-model collaborative optimization. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3597152

[16] Huang, X., Liu, X., Wan, Y., Zheng, Z., Zhang, B., Wang, Y., ... & Zhang, Y. (2025). MVSR3D: An End-to-End Framework for Semantic 3D Reconstruction Using Multi-View Satellite Imagery. IEEE Transactions on Geoscience and Remote Sensing. https://doi.org/10.1109/TGRS.2025.3563498

[17] Gao, M., Chen, F., Wang, L., Zhao, H., & Yu, B. (2024). Swin Transformer-Based Multiscale Attention Model for Landslide Extraction From Large-Scale Area. IEEE Transactions on Geoscience and Remote Sensing,62, 1-14. https://doi.org/10.1109/TGRS.2024.3477910

[18] He, F., Liu, S., Liu, S., Jin, Y., Xie, H., & Tong, X. (2025). HDRoad: An encoder-decoder architecture with hybrid attention and directional prior for efficient road extraction from remote sensing images. ISPRS Journal of Photogrammetry and Remote Sensing,227, 251-264. https://doi.org/10.1016/j.isprsjprs.2025.06.014

[19] Cong, M., Cui, J., Chen, S., Wang, Y., Han, L., Xi, J., ... & Deng, H. (2022). Enhanced shuffle attention network based on visual working mechanism for high-resolution remote sensing image classification. Geocarto International,37(27), 18731-18766. https://doi.org/10.1080/10106049.2022.2143912

[20] Mao, G., Liang, H., Yao, Y., Wang, L., & Zhang, H. (2023). Split-and-shuffle detector for real-time traffic object detection in aerial image. IEEE Internet of Things Journal,11(8), 13312-13326. https://doi.org/10.1109/JIOT.2023.3334742

[21] Ramana, K., Srivastava, G., Kumar, M. R., Gadekallu, T. R., Lin, J. C. W., Alazab, M., & Iwendi, C. (2023). A vision transformer approach for traffic congestion prediction in urban areas. IEEE Transactions on Intelligent Transportation Systems,24(4), 3922-3934. https://doi.org/10.1109/TITS.2022.3233801

[22] Yang, Z. X., You, Z. H., Chen, S. B., Tang, J., & Luo, B. (2023). Semisupervised edge-aware road extraction via cross teaching between CNN and transformer. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,16, 8353-8362. https://doi.org/10.1109/JSTARS.2023.3310612

[23] Dong, X., Shi, P., Liang, T., & Yang, A. (2025). CTAFFNet: CNN–transformer adaptive feature fusion object detection algorithm for complex traffic scenarios. Transportation Research Record, 2679(1), 1947-1965. https://doi.org/10.1177/03611981241258753

[24] Xu, F., Zhang, X., Jin, X., Hu, T., & Lu, S. (2025, November). Multi-Source Information Fusion for Recognition of Typical Paved-Road Features. In2025 12th International Conference on Dependable Systems and Their Applications (DSA) (pp. 394-405). IEEE. https://doi.org/10.1109/DSA66321.2025.00053

[25] Wang, Y., Hong, D., Sha, J., Gao, L., Liu, L., Zhang, Y., & Rong, X. (2022). Spectral–spatial–temporal transformers for hyperspectral image change detection. IEEE Transactions on Geoscience and Remote Sensing,60, 1-14. https://doi.org/10.1109/TGRS.2022.3203075

[26] Meng, Q., Zhou, D., Zhang, X., Yang, Z., & Chen, Z. (2024). Road extraction from remote sensing images via channel attention and multilayer axial transformer. IEEE Geoscience and Remote Sensing Letters,21, 1-5. https://doi.org/10.1109/LGRS.2024.3379502

[27] Huang, B., Lu, Y., Yang, R., Tao, Y., Wang, S., & Shi, Y. (2025). HSN-Net: A hybrid segmentation neural network for high-resolution road extraction. IEEE Geoscience and Remote Sensing Letters. https://doi.org/10.1109/LGRS.2025.3558511

[28] Zhang, H., Li, P., Liu, X., Yang, X., & An, L. (2023). An Iterative Semi-supervised Approach with Pixel-wise Contrastive Loss for Road Extraction in Aerial Images. ACM Transactions on Multimedia Computing, Communications and Applications,20(3), 1-21. https://doi.org/10.1145/3606374

[29] Pu, B., Liu, J., Kang, Y., Chen, J., & Yu, P. S. (2022). MVSTT: A multiview spatial-temporal transformer network for traffic-flow forecasting. IEEE transactions on cybernetics,54(3), 1582-1595. https://doi.org/10.1109/TCYB.2022.3223918

[30] Li, W., Hsu, C. Y., Wang, S., Gu, Z., Yang, Y., Rogers, B. M., & Liljedahl, A. (2025). A multi-scale vision transformer-based multimodal GeoAI model for mapping Arctic permafrost thaw. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. https://doi.org/10.1109/JSTARS.2025.3564310