A Comparative Study of Deep Learning Methods for Automated Road Network Extraction from High-Spatial-Resolution Remotely Sensed Imagery

Haochen Zhou, Hongjie He, Linlin Xu, Lingfei Ma, Dedong Zhang, Nan Chen, Michael A. Chapman, and Jonathan Li

Abstract

Road network data are crucial for various applications, such as road network planning, traffic control, map navigation, autonomous driving, and smart city construction. Automated road network extraction from high-spatial-resolution remotely sensed imagery has shown promise in road network data construction. In recent years, the advent of deep learning algorithms has pushed road network extraction towards automation, achieving very high accuracy. However, the latest deep learning models are often less applied in the field of road network extraction and lack comparative experiments for guidance. Therefore, this research selected three recent deep learning algorithms, including dense prediction transformer (DPT), SegFormer, SEgmentation TRansformer (SETR), and the classic model fully convolutional network-8s (FCN-8s) for a comparative study. Additionally, this research paper compares three different decoder structures within the SETR model (SETR_naive, SETR_mla, SETR_pup) to investigate the effect of different decoders on the road network extraction task. The experiment is conducted on three commonly used datasets: the DeepGlobe Dataset, the Massachusetts Dataset, and Road Datasets in Complex Mountain Environments (RDCME). The DPT model outperforms other models on the Massachusetts dataset with superior reliability, achieving a high accuracy of 96.31% and excelling with a precision of 81.78% and recall of 32.50%, leading to an F_1 score of 46.51%. While SegFormer has a slightly higher F_1 score, DPT's precision is particularly valuable for minimizing false positives, making it the most balanced and reliable choice. Similarly, for the DeepGlobe Dataset, DPT achieves an accuracy of 96.76%, precision of 66.12%, recall of 41.37%, and F_1 score of 50.89%, and for RDCME, DPT achieves an accuracy of 98.94%, precision of 99.07%, recall of 99.84%, and F_1 score of 99.46%, confirming its consistent performance across datasets. This paper provides valuable guidance for future studies on road network extraction techniques using deep learning algorithms.

Introduction

Road network data are crucial for various applications, such as road network planning, traffic control, map navigation, autonomous driving, and smart city construction (Senthilnath *et al.* 2020; Wang *et al.* 2021; Zhang *et al.* 2020; Zhou *et al.* 2020; Yang *et al.* 2020; Chen, Zhong, *et al.* 2021; Tan *et al.* 2020). Specifically, by providing detailed information about existing road layouts, these data help urban planners design more efficient and connected cities. Furthermore, road network data are fundamental for the operation of autonomous vehicles. These data, combined with sensor inputs, enhance the safety and efficiency of autonomous driving by providing a comprehensive view of the road environment (Bagloee *et al.* 2016). For these applications, remote sensing offers a promising method to extract road network datasets (Abdollahi *et al.* 2020).

Regarding the methodologies used for this task, they are akin to those used in extracting information from remote sensing imagery: they can be categorized into (1) visual interpretation and manual editing, (2) feature engineering–based traditional machine learning methods, and (3) deep learning–based methods (Chen *et al.* 2022). Visual interpretation and manual editing are both time-consuming and labor-intensive. Governments invest significantly each year to maintain current road network data. The feature engineering–based traditional machine learning method involves manually constructing a feature set based on expert knowledge, such as object length and aspect ratio, and then classifying using traditional machine learning methods. However, this method requires extensive domain knowledge for the development of accurate and comprehensive rules for road network extraction and is prone to overextraction and susceptibility to occlusions and shadows (Chen *et al.* 2022; Yu *et al.* 2021; Wang *et al.* 2016). The advent of deep learning has enabled autonomous road network extraction. Deep learning–based methods learn features directly from data, achieving higher levels of accuracy and speed (Abdollahi *et al.* 2020). Despite their reliance on significant computational resources, advancements in computing technology are addressing this challenge. Consequently, the deep learning–based method that has dominated in computer vision tasks has been successfully applied in remote sensing and road network extraction (Abdollahi *et al.* 2020).

The first significant advancement in road network extraction over the past decade was introduced by Yuan *et al.* (2011). They developed a method using locally excitatory globally inhibitory oscillator networks to cluster well-aligned points representing the extracted roads, demonstrating consistent performance in the road network extraction task. Following this, Bastani *et al.* (2018) unveiled RoadTracer, an innovative approach that combines two convolutional neural network (CNN) models. One model is tasked with identifying the road classification

Haochen Zhou, Hongjie He, and Jonathan Li are with the Department of Geography and Environmental Management, University of Waterloo, Waterloo, ON N2L 3G1, Canada (hongjie.he@uwaterloo.ca; junli@uwaterloo.ca).

Linlin Xu is with the Department of Geomatics Engineering, University of Calgary, Calgary, AB T2N 1N4, Canada.

Lingfe Ma is with the School of Statistics and Mathematics, Central University of Finance and Economics, Beijing 102206, China.

Dedong Zhang and Jonathan Li are with the Department of Systems Design Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada.

Nan Chen is with the School of Computer Science, Xi'an Aeronautical University, Xi'an, SX 710077, China; and the Key Laboratory of Smart Earth, Beijing 100029, China.

Michael A. Chapman is with the Department. of Civil Engineering, Ryerson University, Toronto, ON M5B 2K3, Canada.

Corresponding author: Hongjie He (h69he@uwaterloo.ca) and Jonathan Li (junli@uwaterloo.ca)

Received September 12, 2024, accepted December 8, 2024.

Photogrammetric Engineering & Remote Sensing
Vol. 91, No. 3, March 2025, pp. 163–174.
0099-1112/22/163–174

doi: 10.14358/PERS.24-00100R2

of a given pixel, while the other aids in constructing the road network map. This method has enabled automatic road network extraction, yielding faster and more accurate results than traditional methods. More recently, Pan *et al.* (2021) trained a fully convolutional network (FCN) model using OpenStreetMap (OSM) data. This approach is both generic and automatic, facilitating the extraction of road networks from very-high-resolution (VHR) remote sensing images. The training and validation of the model leverage road centerlines from OSM, significantly reducing the cost and effort required for data labeling.

Although deep learning methods, in conjunction with VHR remotely sensed imagery, have shown promise for road network extraction tasks (Abdollahi *et al.* 2020), it is important to note that these algorithms were primarily designed for semantic segmentation tasks in computer vision (Wu *et al.* 2023; Wang *et al.* 2024). Additionally, their applications in the field of road network extraction are less explored, and there is a lack of comparative experiments to provide guidance. This paper thus conducts a comparative study, focusing on evaluating the efficacy of three state-of-the-art methods—dense prediction transformer (DPT) (Ranftl *et al.* 2021), SegFormer (Xie *et al.* 2021), and SEgmentation TRansformer (SETR) (Zheng *et al.* 2021)—alongside the classical fully convolutional network-8s (FCN-8s) model in road network extraction tasks. Additionally, this paper examines three different variants of SETR to assess the effect of varying decoder designs on the performance of road network extraction. This paper will provide certain guidance for future research on road network extraction.

The rest of the paper is constructed as follows: First, we discuss the related work in the field of road network extraction and comparative studies. The next section introduces the methodology used in this paper, including model selection, and evaluation metrics. In the next section, we present the experimental results and analysis. The final section draws conclusions and offers suggestions for future studies.

Related Works

This section provides an overview of publicly accessible datasets for road network extraction. We review state-of-the-art (SOTA) segmentation models and the segmentation method applied to remote sensing images, then trace the evolution of road network extraction techniques, ranging from manual annotations and traditional machine learning methods based on feature engineering to contemporary approaches using deep learning. We then briefly examine the related existing methods within these categories, followed by a review of several pertinent comparative studies.

Existing Datasets

There are five commonly used and publicly available datasets for road network extraction from remote sensing images. They are the Massachusetts road dataset (Mnih 2013), the DeepGlobe road network extraction dataset (Demir et al. 2018), the Large Road Segmentation Dataset from Optical Remote Sensing Images of New York (LRSNY) dataset (Chen, Wang, et al. 2021), the SpaceNet road dataset (Van Etten et al. 2018) and Road Datasets in Complex Mountain Environments (RDCME) (Zhang et al. 2022). Table 1 shows the details of each dataset, including the number of images, image size, and pixel size. The Massachusetts Roads Dataset comprises 1171 RGB images of Massachusetts roads, each with a resolution of 1 m/pixel and a size of 1500 1500 pixels. This dataset is segmented into training (819 images), validation (175 images), and test sets (176 images), encompassing a diverse array of environments such as urban, suburban, and rural areas. The DeepGlobe Dataset, originating from the 2018 DeepGlobe Challenge, includes RGB satellite images from Thailand, Indonesia, and India. These images feature a resolution of 0.5 m/pixel and are 1024×1024 in image size. The dataset is structured into 4358 training, 933 validation, and 935 testing images. The LRSNY dataset, situated in central New York City, provides RGB images in two different image sizes: 1000×1000 and 256×256 pixels, each at a resolution of 0.5 m per pixel. It comprises 716 training images, 220 validation images, and 432 test images. The SpaceNet Road Dataset captures VHR RGB images from cities like Las Vegas, Paris, Shanghai, and Khartoum using WorldView-3, with each image having a spatial resolution of 0.3 m per pixel and a size of 3000×3000. This dataset includes 1659 training, 290 validation, and 568 testing images. RDCME is a multispectral image dataset collected from the northwest region of China. The images in this dataset have a resolution of 0.61 m/pixel. The original RDCME dataset contains 12 large-size images. The dataset is cropped into 512×512 images and images that contain no road are removed. There are 286 images for training, 61 images for validation, and 62 images for testing.

Table 1. Existing dataset for road extraction.

Dataset		No. images	Image size (pixels)	Pixel size (m/pixel)
Massachusetts road dataset	Training	1108	1500×1500	1.0
	Validation	14	1500×1500	1.0
	Test	49	1500×1500	1.0
DeepGlobe road network extraction dataset	Training	6226	1024×1024	0.5
	Validation	1243	1024×1024	0.5
	Test	1001	1024×1024	0.5
LRSNY dataset	Training	716	1000×1000	0.5
	Validation	220	1000×1000	0.5
	Test	432	1000×1000	0.5
SpaceNet road dataset	Training	1659	3000×3000	0.3
	Validation	290	3000×3000	0.3
	Test	568	3000×3000	0.3
RDCME dataset	Training	286	512×512	0.61
	Validation	61	512×512	0.61
	Test	62	512×512	0.61

LRSNY = Large Road Segmentation Dataset from Optical Remote Sensing Images of New York; RDCME = Road Datasets in Complex Mountain Environments.

In the ensuing comparative study, the Massachusetts roads and DeepGlobe datasets and RDCME were selected because of their inclusion of various environments and coverage across different countries, respectively. All datasets feature high resolution and have been extensively used in prior research and studies.

Review of State-of-the-Art Segmentation Methods

Recent advancements in segmentation methodologies have introduced state-of-the-art models that push the boundaries of accuracy and adaptability. OneFormer (Jain et al. 2023), a transformer-based framework, unifies semantic, instance, and panoptic segmentation within a single architecture, demonstrating exceptional performance across diverse datasets through its task-conditioned training strategy. Similarly, MedSegDiff-V2 (Wu et al. 2023) integrates diffusion probabilistic models with transformers to improve medical image segmentation, using novel conditioning techniques that enhance feature representation and accuracy. Complementing these is Mamba-UNet (Wang et al. 2024), a UNet-like architecture enriched with attention mechanisms to capture local and global features, making it highly effective for detailed segmentation tasks. These innovations reflect the growing trend of leveraging transformer-based architectures and attention mechanisms to achieve precise and robust segmentation across various domains.

Image Segmentation for Remote Sensing Datasets

In parallel, segmentation efforts tailored to remote sensing datasets have underscored the importance of dataset quality and preprocessing techniques. Subedi et al. (2023) demonstrated the utility of high-resolution National Agriculture Imagery Program (NAIP) imagery for large-scale land use mapping, emphasizing its relevance for road extraction tasks where data resolution is critical. Marchand et al. (2023) explored 3D surface mesh reconstruction from remote sensing data, offering insights applicable to 3D road network extraction. Additionally, Mezouar et al. (2023) proposed a K-Means-based orthorectification algorithm to correct geometric distortions in satellite images, ensuring spatial accuracy essential for reliable segmentation. Together, these studies highlight the potential of integrating preprocessing strategies and advanced segmentation techniques to maximize the utility of remote sensing datasets for tasks like road extraction.

As the road extraction task can be treated as a binary segmentation task, the segmentation method can be applied to road extraction tasks.

Road Network Extraction from Remotely Sensed Imagery

The initial road extraction approach involved manual annotation, which includes visual interpretation followed by the generation of polygons. As for feature engineering–based machine learning, features such as object length and aspect ratio are meticulously crafted through expert knowledge, and traditional machine learning algorithms are used for classification. These methods were the dominant paradigm before the advent of methods based on deep learning algorithms for road network extraction (Chen *et al.* 2022).

Traditional Feature Engineering–Based Machine Learning Methods

Traditional feature engineering–based machine learning methods can be classified into two categories: handcrafted feature–based methods and morphological feature–based methods (Chen *et al.* 2022). These methods use distinct techniques for feature extraction.

Morphological features are extensively used in road network extraction because of their uniform shape and consistent appearance, as noted by Alshehhi and Marpu (2017). The domain has witnessed significant advancements in morphological operations, with a succession of studies building upon the foundational insights of their predecessors to refine techniques and address emerging challenges. Initially, the primary methods relied on quantifiable metrics such as widths and shapes, using techniques including binarization, expansion, erosion, opening, and closing. These early techniques were prone to inducing shape biases, a limitation that was subsequently addressed by Valero *et al.* (2010) through the introduction of advanced directional morphological operators known as path openings and path closings. This pivotal development recognized the directional properties of road patterns and influenced further research, such as the work of Chaudhuri *et al.* (2012), who crafted operators specifically designed to harness the directional and morphological attributes of roads, enhancing the accuracy of the extraction process.

Subsequent innovations involved the integration of both low-level and high-level processing techniques, as exemplified by Bae *et al.* (2015). Their approach used various attributes including widths, contrast, orientations, lengths, and sophisticated classifiers like graph cutting. Concurrently, Leninisha and Vani (2015) introduced a geometric active deformable model predicated on width and color, designed to adapt dynamically to fluctuating road conditions. Expanding upon these morphological principles, Courtrai and Lefèvre (2016) implemented morphological path filters on regions rather than mere pixels, which augmented the granularity of processing and improved road detection within intricate environments. Similarly, Grinias *et al.* (2016) adopted an unsupervised approach that amalgamated geometric features with statistical models, notably the Markov random fields and random forest methods. Building on these cumulative advances, Zang *et al.* (2017) introduced a pixel value–based enhancement technique, further refining the precision and clarity of road imagery analysis. Collectively, these developments illustrate a progression toward increasingly sophisticated and integrated road network extraction methodologies, effectively combining geometry, statistical modeling, and advanced image processing techniques.

While morphological feature–based approaches can successfully extract road shape features, they are typically susceptible to light and contrast variations, occlusions, and other disturbances (Chen *et al.* 2022; Wang *et al.* 2016). Handcrafted features, on the other hand, can meet these special constraint texture features. Following the extraction of features, classifiers are applied to make final classifications. Classical classifiers include support vector machines, decision trees, Hough forests, tensor voting, and others (Chen *et al.* 2022; Oussama *et al.* 2023). Handcrafted features have achieved much success in past decades (Krylov and Nelson 2014; Poullis and You 2010; Wegner *et al.* 2015).

Deep Learning–Based Methods

Recently, the development of deep learning has enabled significant advancements in various computer vision–related fields, particularly in road network extraction from remote sensing images (Kestur *et al.* 2018; Zhang *et al.* 2018). These advancements have manifested across four key methods: patch-based CNN models, generative adversarial networks (GANs), encoder-decoder networks, and feature fusion.

Patch-based CNN models train on small image patches and use a sliding patch mechanism to predict road networks. This method has been refined by integrating CNN features with low-level road characteristics to enhance mapping accuracy (Alshehhi *et al.* 2017). Further optimizations in CNN architecture have been specifically tailored for road mapping, enhancing precision and detailed road structure representation (Chen, Wang *et al.* 2021; Saito and Aoki 2015; Li *et al.* 2016). GANs, conceptualized by Ian Goodfellow *et al.* in 2014, include a generator that creates realistic samples and a discriminator that evaluates their authenticity. This approach has been adapted to identify challenging road segments hidden by shadows or occlusions (Zhang *et al.* 2019). Integrating GANs with other architectures like U-Net and FCN has improved road segmentation accuracy, despite challenges like gradient instability and complex training processes (Shamsolmoali *et al.* 2021; Senthilnath *et al.* 2020). Encoder-decoder networks use an encoder to distill essential features from images and a decoder to reconstruct spatial dimensions, maintaining high-resolution and detailed image reconstruction. This structure has led to innovations such as direction-aware and context-sensitive models that improve the understanding of linear road features (Xu *et al.* 2021). Feature fusion techniques merge multiple feature maps to enrich semantic information, enhancing the accuracy of image segmentation. This approach leverages features at various scales and combines them to capture a comprehensive representation of the scene. Innovations in this area include integrating Atrous Spatial Pyramid Pooling with residual networks to achieve nuanced multi-scale feature representation, which enhances detail preservation and context sensitivity in segmentations (Tan *et al.* 2021; Zhang *et al.* 2022; Ma *et al.* 2020; Wu *et al.* 2021).

Although deep learning–based methods, especially those using patch-based CNNs and GANs, have grown rapidly and shown promising results in the field of road extraction, comparative studies specifically focusing on encoder-decoder design networks remain limited. Therefore, in this research, we selected DPT, SegFormer, and SETR, as they are all based on encoder-decoder designs, for the comparative study.

Existing Comparative Studies

There are many comparative studies focused on the application of deep learning approaches in remote sensing, and the methodology and evaluation metrics could be used in this paper to evaluate the experimental results.

Ye *et al.* (2020) analyzed the accuracy of area-based dense image-matching techniques at subpixel levels for remote sensing applications. They evaluated 12 algorithms using correlation-based similarity measures and subpixel estimation techniques across simulated and real-world datasets. The findings highlighted performance differences among these algorithms and suggested aligning algorithm selection with specific application requirements and challenges posed by aliasing effects. Cai *et al.* (2021) compared the effectiveness of deep learning models—FCN, U-Net, and DeepLabv3+—for detecting rooftops in aerial photographs using a high-quality dataset from Kitchener-Waterloo, Ontario. The research examined these models with different dataset volumes and evaluated them based on metrics like Intersection over Union (IoU) and F_1 score. The study found that DeepLabv3+ showed the highest accuracy, emphasizing the importance of choosing appropriate deep-learning algorithms and loss functions for urban planning. Xu *et al.* (2023) compared 12 different loss functions for road segmentation in remote sensing images using the D-LinkNet architecture. They used the Massachusetts roads dataset and the DeepGlobe road extraction dataset, evaluating each loss function's effectiveness with metrics like precision, recall, and IoU. The findings suggested that region-based and compound loss functions, particularly Focal Tversky loss and Lovasz-Softmax loss, outperformed distribution-based counterparts, recommending further exploration into loss function combinations for improved road segmentation accuracy. Kumbasar *et al.* (2023) compared 12 CNN models, including U-Net, Feature Pyramid Network (FPN), LinkNet, SegNet, FCN, and six residual U-Net variants, for

Table 2. Comparison matrix of selected models.

Model	Architecture	Key Features	Parameters
FCN	Fully convolutional network	Skip connections for combining coarse and fine features	~23.52M
DPT	Vision Transformer + Convolutional Decoder	Global receptive field in all layers, convolutional decoder	~86M
SETR	Pure Transformer Encoder + Custom Decoder	Sequence-to-sequence transformer; maintains full resolution	~318M
SegFormer	Hierarchical Transformer + MLP Decoder	Positional encoding–free, multi-scale feature generation	~64M

DPT = dense prediction transformer; FCN = fully convolutional network; SETR = SEgmentation Transformer; MLP = multilayer perceptron.

building segmentation using satellite and UAV imagery. Models were trained on the Inria Aerial Image Labeling dataset and evaluated on three datasets with diverse characteristics: Inria, Massachusetts Buildings, and Syedra Archaeological Site. Residual U-Net models outperformed others, demonstrating the effectiveness of residual connections in preserving spatial details. LinkNet with EfficientNet-B5 excelled on archaeological site data, showcasing strong generalization to unique features. The analysis highlights the significance of architectural design and residual blocks for accurate and versatile segmentation across varying datasets.

The most closely related research is the work completed by Xu et al (2023); however, existing studies often lack analysis of the latest deep learning algorithms. Thus, this research aims to address this gap by conducting experiments that include recent advancements in deep learning, thereby contributing valuable findings to the field.

Methodology

The goal of this study is to discover and compare the effectiveness of different segmentation models in extracting road networks from remote sensing images. The first part of the study examines FCN-8s, a renowned model in semantic segmentation that serves as a foundational benchmark in this field. Since its inception, FCN-8s has laid the groundwork for future advancements in segmentation techniques. However, as the field has evolved, newer models have been developed that address some of the challenges and limitations associated with FCN-8s.

The original FCN algorithm has been studied and referred to a lot, which has given us a lot of information about what it can and cannot accomplish in road network extraction tasks. To examine the evolution of the field and address historical challenges, we will review models from various years that demonstrate the growth and transformation of segmentation technology. Besides the FCN-8s model, this paper also focuses on other state-of-the-art models such as DPT, SegFormer, and SETR. All of them are the new models introduced in recent years, and they perform well on semantic segmentation. Therefore, they were selected for the comparative study to explore their performance on road network extraction. Table 2 shows the comparative matrix of these models. Additionally, all three variants of SETR—SETR_naive, SETR_mla, and SETR_pup—are tested to assess the effect of varying decoder designs on the performance of road network extraction.

Fully Convolutional Network-8s

It was Long *et al.* (2015) who made significant advances in the field of semantic segmentation by introducing FCN-8s. Traditional CNNs perform well at classifying objects, but they are not particularly adept at making predictions pixel by pixel. FCN-8s addressed this by converting every fully connected layer to a convolutional layer. This adaptation allowed the network to handle inputs of any size and produce outputs that correspond spatially to the inputs. This design enabled continuous training and inference, which enhanced its flexibility and utility for a wide range of image sizes.

Upsampling layers and skip connections allow FCN-8s to combine detailed data from shallow layers with more general features. This was one of its most significant innovations. This method greatly enhanced the segmentation quality, enabling the model to display both the general structure and fine details of roads. When tested on standard datasets such as PASCAL Visual Object Classes, FCN-8s performed exceptionally well, demonstrating substantial improvements over previous segmentation methods. For instance, the FCN-8s variant improved border delineation and interior area accuracy considerably, leading to more precise road maps.

FCN-8s, through a strategy of layer-by-layer refinement, first upsamples the deepest feature map by a factor of 2, then merges it with the feature map from the previous layer. After another 2× upsampling, it continues by merging with an even earlier layer's feature map, ultimately achieving an 8× upsampling to match the size of the input image. This strategy helps to refine boundaries and enhance the segmentation performance for small objects (Figure 1, see next page).

Dense Prediction Transformer

A big step forward has been made with the DPT in dense prediction tasks like semantic segmentation and monocular depth predictions. DPT uses an encoder-decoder structure and is built on a vision transformer framework. This makes it better at handling complex image analysis jobs. Its main innovation is that it keeps the spatial resolution the same throughout all its steps of processing, making sure that there is always a global receptive field. This approach enables DPT to generate predictions that not only are more detailed but also exhibit superior global consistency compared with traditional FCNs (Ranftl *et al.* 2021).

The architecture of DPT excels when trained on extensive databases. It establishes new benchmarks in tasks such as semantic segmentation and depth estimation. It also demonstrates adaptability and performs effectively on smaller datasets. Because the model captures fine details while also comprehending the broader context, it is well suited for tasks like urban planning and navigation, which require precise and comprehensive image analysis. Overall, the introduction of DPT represents a milestone in the field, enabling enhanced accuracy and understanding in image-based predictions (Ranftl *et al.* 2021).

Figures 2 and 3 show the structure of DPT. Figure 2 shows the overview of architecture. The input image is transformed into tokens by extracting nonoverlapping patches followed by a linear projection of their flattened representation. The image embedding is augmented with a positional embedding and a patch-independent readout token is added. The tokens are passed through multiple transformer stages. We reassemble tokens from different stages into an image-like representation at multiple resolutions. Fusion modules progressively fuse and upsample the representations to generate a fine-grained prediction. Figure 3 shows the details of DPT's components. Figure 3a shows the

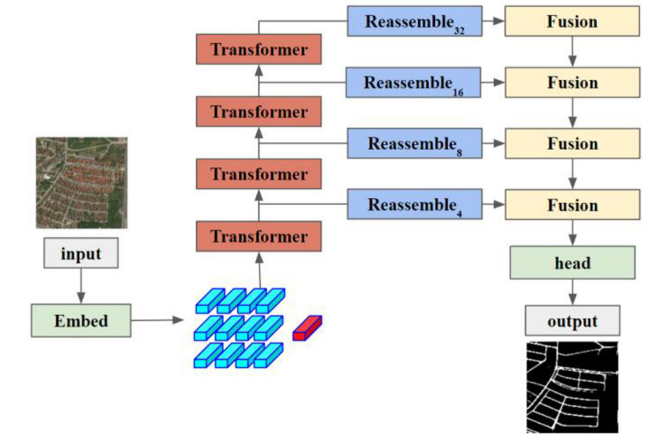


Figure 2. Dense prediction transformer main structure that includes convolutional layers (modified from Ranftl *et al.* 2021).

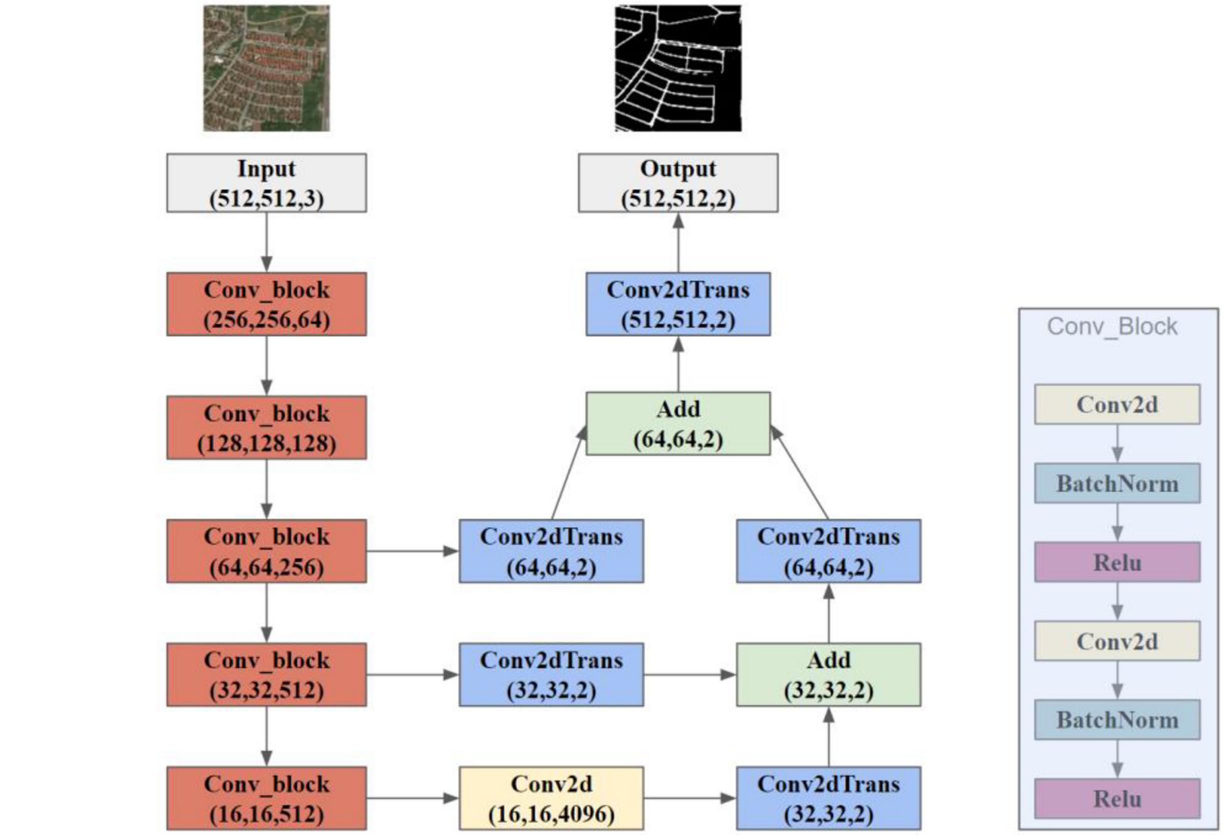


Figure 1. Fully convolutional network-8s main structure that includes convolutional layers (modified from Long *et al.* 2015).

Reassembles operation, in which tokens are assembled into feature maps with 1/s the spatial resolution of the input image, where s denotes the output size ratio of the recovered representation with respect to the input image and Figure 3b shows Fusion blocks combining features using residual convolutional units and upsampling the feature maps (Ranftl *et al.* 2021).

SegFormer

SegFormer came about because of the need for semantic segmentation models that work well and can be scaled up, especially for high-resolution satellite images (Xie *et al.* 2021). It is a combination of a transformer-based encoder and a simple multilayer perceptron (MLP) decoder that strikes a good balance between accuracy and speed. Its hierarchical transformer encoder picks up features at more than one scale, which makes it good at dealing with the different sizes and shapes of roads in satellite images.

One aspect that distinguishes SegFormer is its design for rapid operation. Traditional transformer models often require substantial computing power when processing high-resolution inputs. SegFormer addresses this issue with a hierarchical structure that manages features of varying sizes, making the computations more efficient while maintaining high accuracy. SegFormer's design renders it an excellent choice for large-scale satellite data analysis, where both accuracy and speed are critical.

In terms of performance, SegFormer has demonstrated its proficiency in various segmentation tasks by establishing new benchmarks for speed and accuracy. Its ability to produce detailed segmentation maps while remaining computationally efficient makes it highly effective for road network extraction, which requires a precise understanding of how roads interconnect. SegFormer's innovative approach represents a significant advancement in the field and provides academics and practitioners with a powerful tool.

Figure 4 shows structure of the SegFormer including two main modules: (1) a hierarchical Transformer encoder to extract coarse and fine features, where "FFN" indicates a feed-forward network which are also shown by Figure 5; and (2) a lightweight All-MLP decoder to directly fuse these multi-level features and predict the semantic segmentation mask.

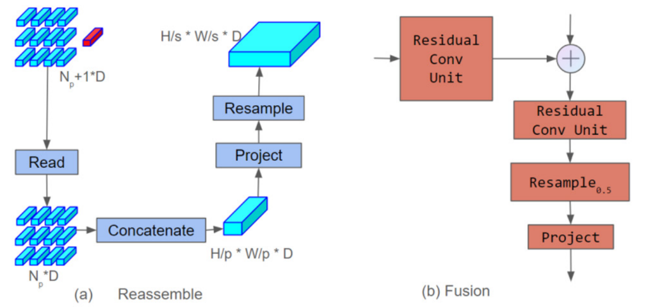


Figure 3. Details of dense prediction transformer components (modified from Ranftl *et al.* 2021).

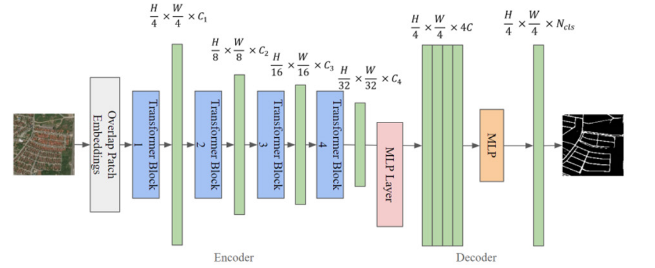

$\frac{H}{4} \times \frac{W}{4} \times C_1$
$\frac{H}{8} \times \frac{W}{8} \times C_2$
$\frac{H}{16} \times \frac{W}{16} \times C_3$
$\frac{H}{32} \times \frac{W}{32} \times C_4$
$\frac{H}{4} \times \frac{W}{4} \times 4C$
$\frac{H}{4} \times \frac{W}{4} \times N_{cls}$
Overlap Patch Embeddings
Transformer Block 1
Transformer Block 2
Transformer Block 3
Transformer Block 4
MLP Layer
MLP
Encoder
Decoder

Figure 4. SegFormer main structure that includes convolutional layers (modified from Xie *et al.* 2021). MLP = multilayer perceptron.

SEgmentation TRansformer

Zheng *et al.* (2021) introduced the SETR model, using a transformer-based design instead of the more commonly used CNNs. Conventional methods predominantly use CNNs to extract features and process local context. SETR reconsiders this approach by treating image segmentation as a sequence-to-sequence prediction challenge, offering a novel perspective. It achieves this by encoding images as groups of patches and then processing these groups with a transformer, ensuring that each layer comprehends the broader context. Unlike CNN-based models, which generally have hierarchical structures and local receptive fields, this method diverges significantly from those characteristics. Based on different kinds of decoders, there are three variants of SETR. A straightforward upsampling method is used with the SETR_naive variant. Subsequently, bilinear upsampling is used to restore the resolution to that of the original image. This is achieved by directly projecting the features obtained from the transformer onto the various image categories. Progressive upsampling is implemented by the SETR_pup encoder, also known as the progressive upsampling encoder, which uses a more intricate method. This is achieved by alternating between upsampling and convolution layers. In each iteration, upsampling is restricted up to 2×. The MLA encoder in SETR_mla, also referred to as multi-level feature aggregation, integrates features from different stages of the generator. SETR_mla ensures uniform quality across all layers, which facilitates the consistent representation of features. Conversely, standard feature pyramid networks possess varying feature sizes.

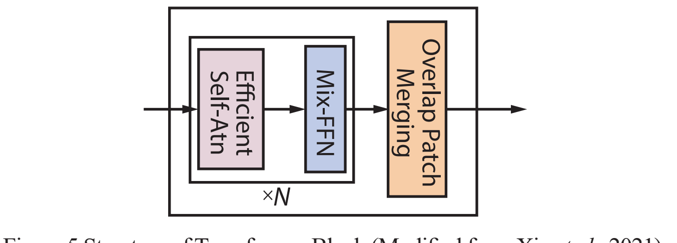


Figure 5 Structure of Transformer Block (Modified from Xie *et al.*, 2021).

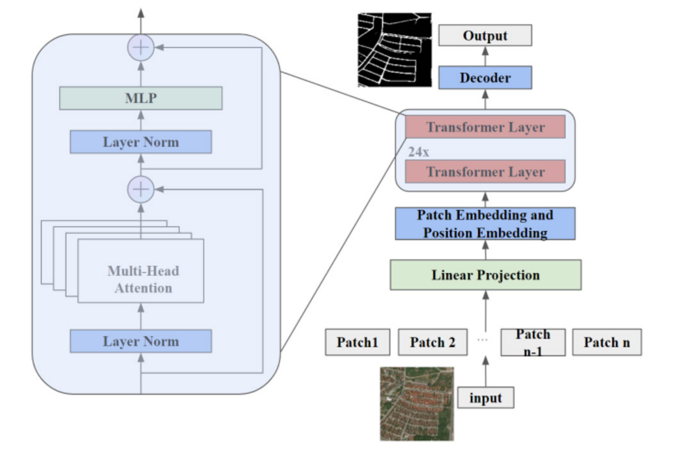


Figure 6. SETR main structure that includes convolutional layers (modified from Zheng *et al.* 2021). MLP = multilayer perceptron; SETR = SEgmentation TRansformer.

Figures 6 and 7 show the structure of SETR. Figure 6 shows the main structure of SETR, where the image is split into fixed-size patches, and each of the patches is linearly embedded, then position embeddings are added and the resulting sequence of vectors is fed to a standard transformer encoder, and then one of the three decoders mentioned above is applied. Figure 6 illustrates two specialized designs of decoders: Figure 7a shows progressive upsampling, which is the key part of SETR_pup, and Figure 7b shows multi-level aggregation, which contributes to SETR_mla (Zheng *et al.* 2021).

Evaluation Metrics

The commonly used evaluation metrics for road segmentation are precision, recall, F_1 score, (pixel) accuracy, mean accuracy (mAcc) and mean IoU (mIoU), which can be calculated as follows:

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{1}$$

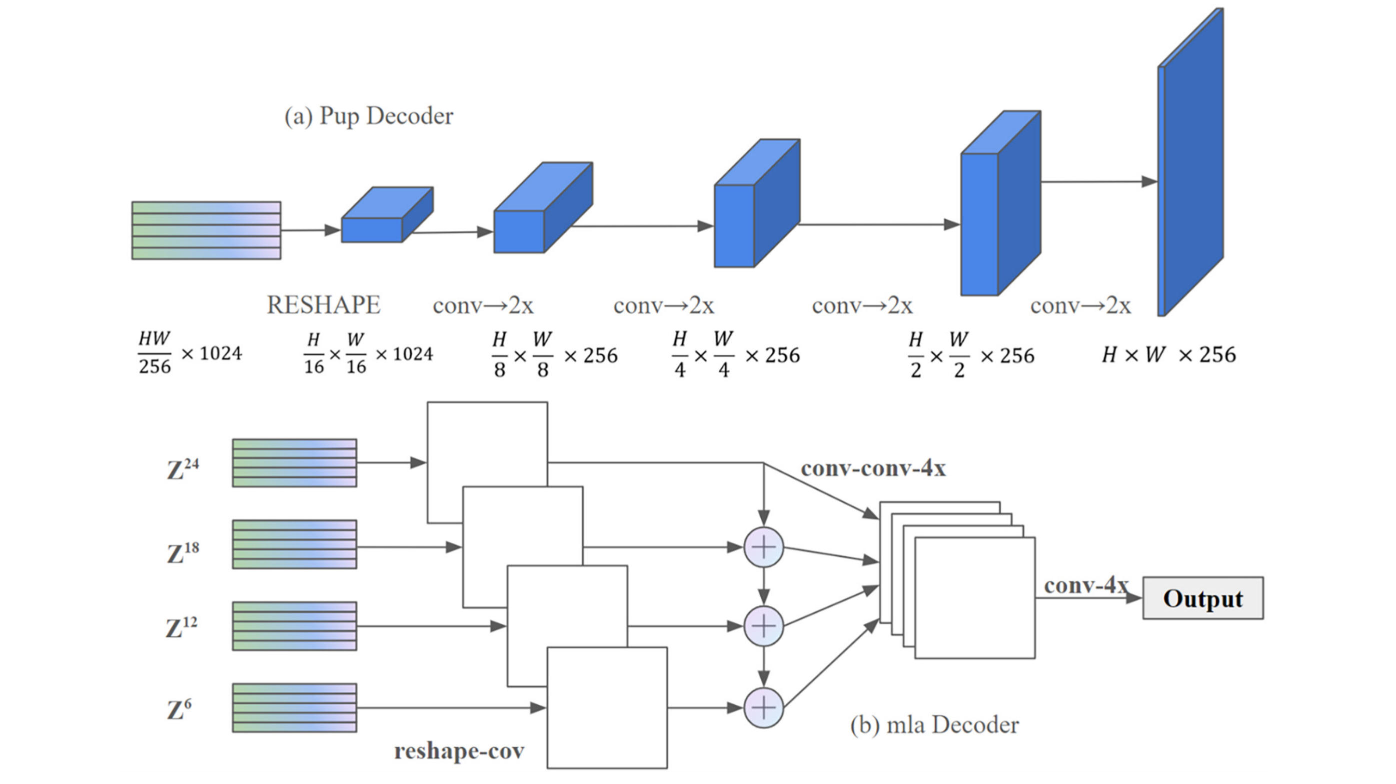

(a) Pup Decoder
RESHAPE
conv→2x
conv→2x
conv→2x
conv→2x
$\frac{HW}{256} \times 1024$
$\frac{H}{16} \times \frac{W}{16} \times 1024$
$\frac{H}{8} \times \frac{W}{8} \times 256$
$\frac{H}{4} \times \frac{W}{4} \times 256$
$\frac{H}{2} \times \frac{W}{2} \times 256$
$H \times W \times 256$
Z^{24}
Z^{18}
Z^{12}
Z^{6}
conv-conv-4x
conv-4x
Output
reshape-cov
(b) mla Decoder

Figure 7. SETR decoder designs (modified from Zheng *et al.* 2021). SETR = SEgmentation TRansformer.

$$Recall = \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FN})} \tag{2}$$

$$F_1 = \frac{2\times\left(Precision \times Recall\right)}{(Precision + Recall)} = \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} \tag{3}$$

$$Accuracy = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}+\mathrm{FP}+\mathrm{TN}} \tag{4}$$

$$mAcc = \frac{Accuracy_1 + Accuracy_2}{2} \tag{5}$$

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} \tag{6}$$

$$\mathrm{mIoU} = \frac{\mathrm{IoU}_1 + \mathrm{IoU}_2}{2} \tag{7}$$

True positive (TP) and false negative (FN) are outcomes in which the model correctly and incorrectly predicts the positive class, respectively. Similarly, FP and FN are outcomes where the model correctly and incorrectly predicts the negative class, respectively. Precision is the number of correctly identified road pixels out of all the pixels that were predicted to be roads. Recall, on the other hand, is the number of correctly identified road pixels out of all the road pixels in the initial image. The F_1 score is the harmonic mean of accuracy and recall (Taha and Hanbury 2015). While accuracy is the number of right predictions out of all the predictions, mIoU is a common way to measure semantic segmentation, which evaluates the areas where the predicted segmentation and the ground truth meet. To find it, the average of the IoU scores for all classes in the dataset is used. This gives a single performance number that can be used to compare models directly (Everingham *et al.* 2009). We also monitor the training loss of the models, which measures the discrepancy between the predicted outputs and the true labels. A decreasing trend in training loss over time indicates effective learning and improved model performance during the training process. It gives a direct measure of the model's error and information about how it learns and how it converges (Goodfellow *et al.* 2016).

Experiment and Results

Data Selection

Three road datasets are used in this work: the Massachusetts roads dataset, the DeepGlobe road extraction dataset, and the RDCME dataset. All three datasets are widely used remote-sensing road datasets (Xu *et al.* 2023; Buslaev *et al.* 2018; He *et al.* 2019). The Massachusetts roads dataset contains aerial RGB images having a resolution of 1 m/pixel and a size of 1500×1500 pixels, separated into three sets, which are 819 images for the training set, 176 images for the validation set, and 176 images for the test set. The DeepGlobe road extraction dataset contains satellite RGB images. There are 4368, 928, and 930 images for the training, validation, and test sets respectively. The images are 0.5 m/pixel and the size of the images is 1024×1024. The RDCME dataset is specifically for mountainous areas, which are slightly different from the previous two: this dataset is cropped to 512×512 size and only the images that contain roads are used; there are 286 images for training, 61 images for validation, and 62 images for testing. Table 3 shows example images and labels of three datasets.

Model Training

The training for this project component was primarily conducted on a Windows-based system equipped with an Nvidia GeForce RTX 3090 GPU and an Intel Core i7 CPU with 32 GB of RAM. The hyperparameters remained consistent across all models during training to facilitate fair comparison. The initial setup involved adjusting model clarity, with all models set to a crop size of 512×512, which is both a default and an optimal performance setting. However, because of memory constraints encountered by some models at this resolution, batch sizes were uniformly reduced to 1 from 4 to maintain consistency across experiments. Stochastic gradient descent served as the optimizer for all models, featuring a learning rate of 0.001, momentum of 0.9, and a weight decay of 0.0005. A polynomial decay schedule was applied where the learning rate began at 0.001 and decreased to a minimum of 1×10^{-4} over 400 000 iterations, with a decay power of 0.9. The training loop of the model was structured around iterations, capped at a maximum of 400 000, with validation intervals every 20 000 iterations.

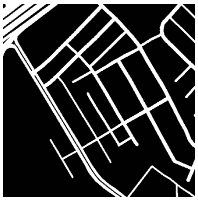
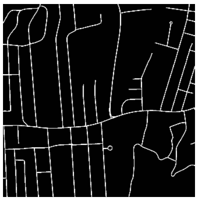
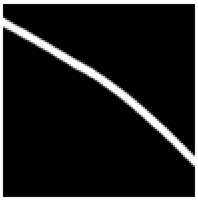
Results

The FCN-8s, DPT, and SegFormer exhibited similar characteristics across three experimental datasets, with their training loss decreasing gradually in a relatively smooth manner. Conversely, the training loss for SETR_naive was relatively unstable, lacking a clear downward progression. Figure 8-11 show the training loss on the DeepGlobe as an example. In fact, Figure 12-14 illustrate that all SETR variants exhibit relatively subtle decreases in loss compared with other models.

The changes in mIoU and mAcc during the validation process were also recorded. FCN-8s shows an overall upward trend, while SegFormer's showed a more noticeable upward trend in the Massachusetts dataset. DPT's mAcc demonstrated a significant upward trend in training across both datasets. Conversely, SETR_naive showed mediocre performance in both datasets, with not very apparent increases in mIoU. Of its two other variants, SETR_mla had a decent upward trend in mIoU in the DeepGlobe dataset, while SETR_pup performed well in all datasets, with mIoU showing significant increases. Figure 15 illustrates how the four models performed on DeepGlobe as an example, while Figure 16 shows the performance of SETR variants.

In the comprehensive analysis of road network extraction capabilities across the three datasets, an evaluative approach was used to assess the performance of four distinct models: FCN-8s, DPT, SegFormer, and SETR_naive, representing the SETR variants, followed by a detailed comparison among the SETR variants themselves—SETR_naive, SETR_mla, and SETR_pup.

Table 3. Examples of three datasets.

	DeepGlobe	Massachusetts	RDCME
Image			
Label			

RDCME = Road Datasets in Complex Mountain Environments.

Table 4 shows the performance analysis on three datasets, illustrating that DPT excels, demonstrating robust adaptability across various tasks. The FCN-8s model, while generally showing decent performance, reveals its limitations in scenarios characterized by sparse road pixels, indicating a challenge in processing less-defined road networks. SegFormer registers moderate performance, reflecting average capabilities that do not particularly excel in any specific aspect. SETR_naive, selected to represent the SETR family in this initial analysis, exhibits the least satisfactory outcomes, indicating it struggles significantly under complex scenarios.

Further exploration of the SETR variants reveals significant differences in their performance. As shown in Table 5, in the DeepGlobe dataset, SETR_mla exhibits only moderate capabilities, better than SETR_naive but not achieving high performance, while SETR_pup stands out with its relative superiority, suggesting that it includes optimizations that ameliorate some of the common deficiencies found in the other variants.

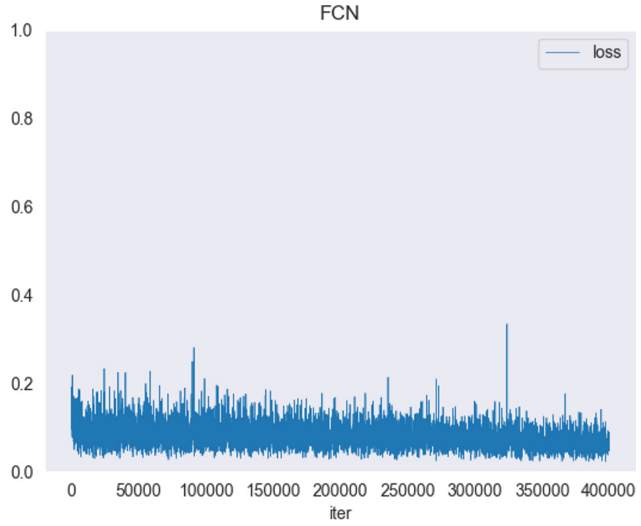


Figure 8. Loss of FCN.

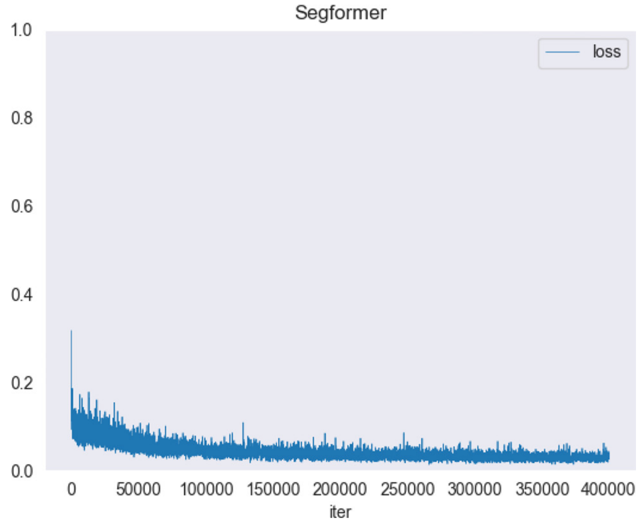


Figure 9. Loss of Segformer.

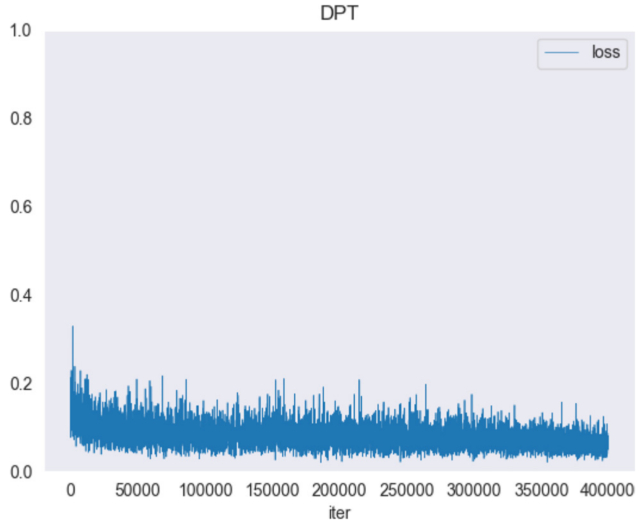


Figure 10. Loss of DPT.

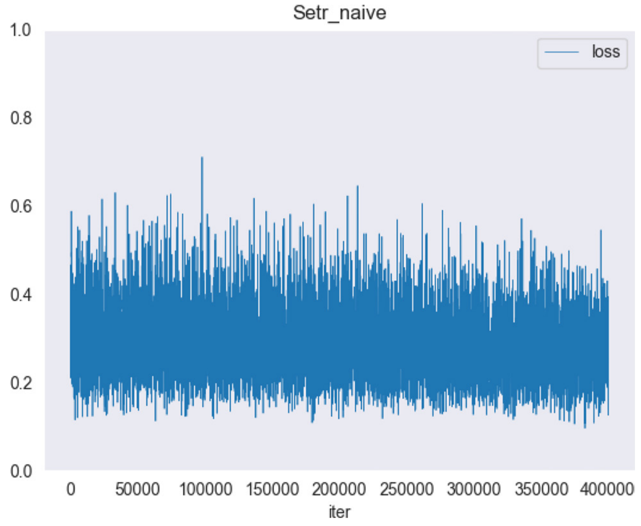


Figure 11. Loss of SETR_naive.

DPT = dense prediction transformer; FCN = fully convolutional network; SETR = SEgmentation TRansformer.

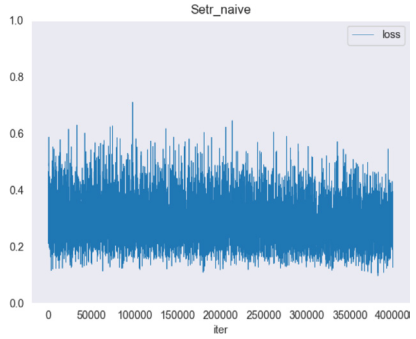


Figure 12. Loss of SETR_naive.

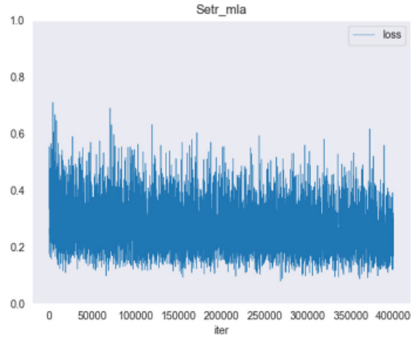


Figure 13. Loss of SETR_mla.

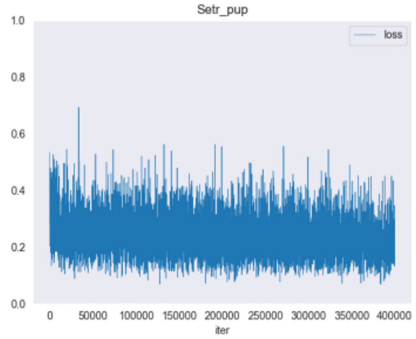


Figure 14. Loss of SETR_pup.

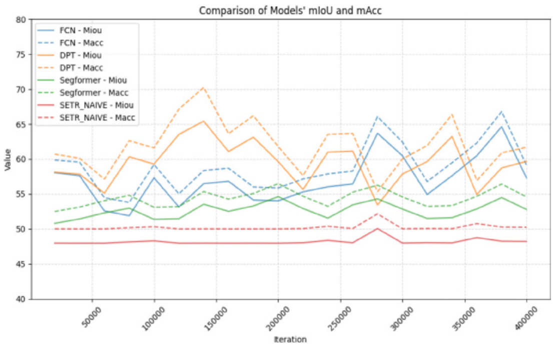


Figure 15. mIoU and mAcc of four models on DeepGlobe. mAcc = mean Accuracy; mIoU = mean Intersection over Union;

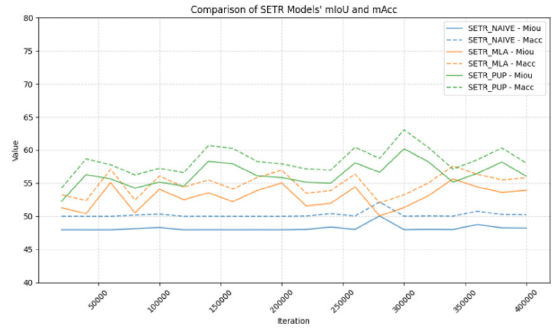
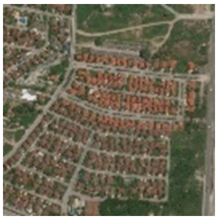
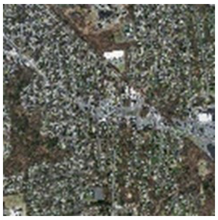
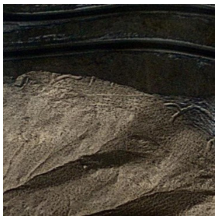
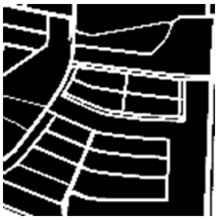
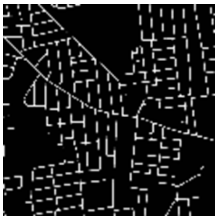
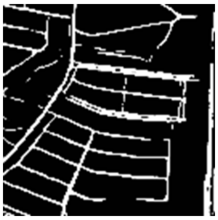
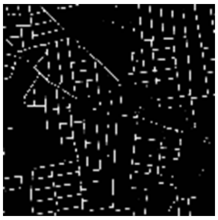
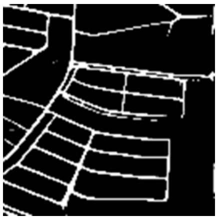
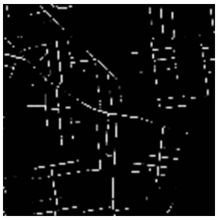
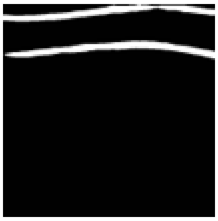
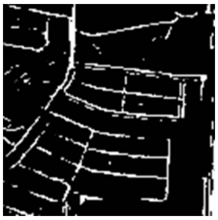
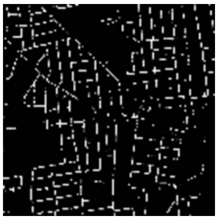
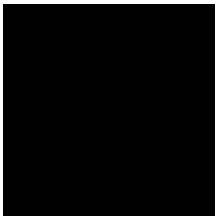
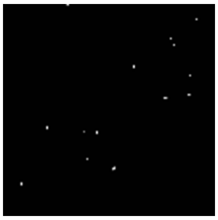


Figure 16. mIoU and mAcc of SETR variants on DeepGlobe. mAcc = mean Accuracy; mIoU = mean Intersection over Union; SETR = SEgmentation TRansformer.

Table 4. Examples of road network extraction results for three datasets using four different models.

	DeepGlobe	Massachusetts	RDCME
Image			
Label			
DPT			
FCN-8s			
SegFormer			
SETR_naive			

DPT = dense prediction transformer; FCN = fully convolutional network; RDCME = Road Datasets in Complex Mountain Environments; SETR = SEgmentation TRansformer.

The performance analysis of four models (FCN-8s, DPT, SegFormer, and SETR_naive) across the DeepGlobe, Massachusetts, and RDCME datasets reveals significant variations in effectiveness.

As shown in Table 6, FCN-8s and DPT consistently exhibit strong performance, particularly on the RDCME dataset, where they achieve near-perfect accuracy, precision, recall, and F_1 scores. However, both models struggle with recall and F_1 scores on the more challenging DeepGlobe and Massachusetts datasets, highlighting a potential limitation in identifying true positives. SegFormer, while achieving high accuracy, underperforms on DeepGlobe and Massachusetts because of its low recall and F_1 scores, suggesting poor sensitivity to positive class predictions. SETR_naive demonstrates the weakest overall performance, particularly on the Massachusetts dataset, where its recall and F_1 scores are negligible, though it performs moderately well on RDCME.

Table 7 shows the evaluation of SETR variants. SETR_pup emerges as the most balanced variant, with noticeable improvements in recall and F_1 scores across datasets, particularly outperforming the SETR_naive and SETR_mla variants on DeepGlobe and Massachusetts. SETR_mla, while achieving high precision, fails to address the recall deficiencies that hinder its F_1 score. All SETR variants perform exceptionally well on RDCME, achieving near-perfect metrics, indicating strong dataset-specific generalization. However, their limited success on DeepGlobe and Massachusetts highlights challenges in adapting to complex or diverse data, emphasizing the need for further optimization in transformer-based segmentation models.

We conducted a generalizability performance experiment by training the model on the DeepGlobe dataset and applying it to the Massachusetts dataset and RDCME. The model achieved a mIoU of 58.60% and a mAcc of 70.35% on the Massachusetts dataset. In contrast, its performance on RDCME was lower, with a mIoU of 49.95% and a mAcc of 52.42%. This discrepancy may be attributed to the similarities in road features between the DeepGlobe and Massachusetts datasets, as both primarily contain urban areas, whereas RDCME predominantly features mountain roads.

Conclusion and Future Recommendations

Based on the analysis of training processes and comparisons among all models, it is evident that the DPT model emerges as the most suitable choice for road network extraction tasks because of its commendable accuracy and consistent performance across different datasets. It achieved 96.76% accuracy, 66.12% precision, 41.37% recall, and a 50.89% F_1 score on the DeepGlobe Dataset. While the FCN-8s is noteworthy for its precision, its recall sometimes falls short of accuracy, indicating that although it is effective, it may not always be reliable under varying circumstances.

Additionally, there are intriguing findings related to the variants of the SETR. Specifically, the performance of SETR_pup closely mirrors that of the best-performing DPT model. However, the experiments reveal that the loss associated with SETR models exhibits instability, which could be linked to the dataset size. This issue becomes more pronounced when transitioning to the Massachusetts dataset, which is smaller in size. Despite these challenges, the SETR variants, particularly SETR_pup, show promise for road network extraction tasks, suggesting that their effectiveness might be further enhanced with larger datasets.

Regarding future research recommendations, the significant influence of dataset size and complexity on model performance necessitates a focus on customizing models for specific types of datasets. A deeper understanding of how different models respond to varying dataset characteristics could facilitate the development of more tailored and effective model designs. For example, using synthetic datasets to train models could enhance understanding of the limits and capabilities of each model in controlled environments. Furthermore, cross-model learnings and hybrid approaches could prove beneficial. Insights drawn from the strengths and weaknesses of each model may inform the development of hybrid models that integrate the high precision of FCN-8s with the balanced performance of DPT and the promising attributes of SETR variants. Such hybrid strategies could potentially leverage the unique features of each model to enhance overall extraction accuracy and robustness.

Acknowledgments

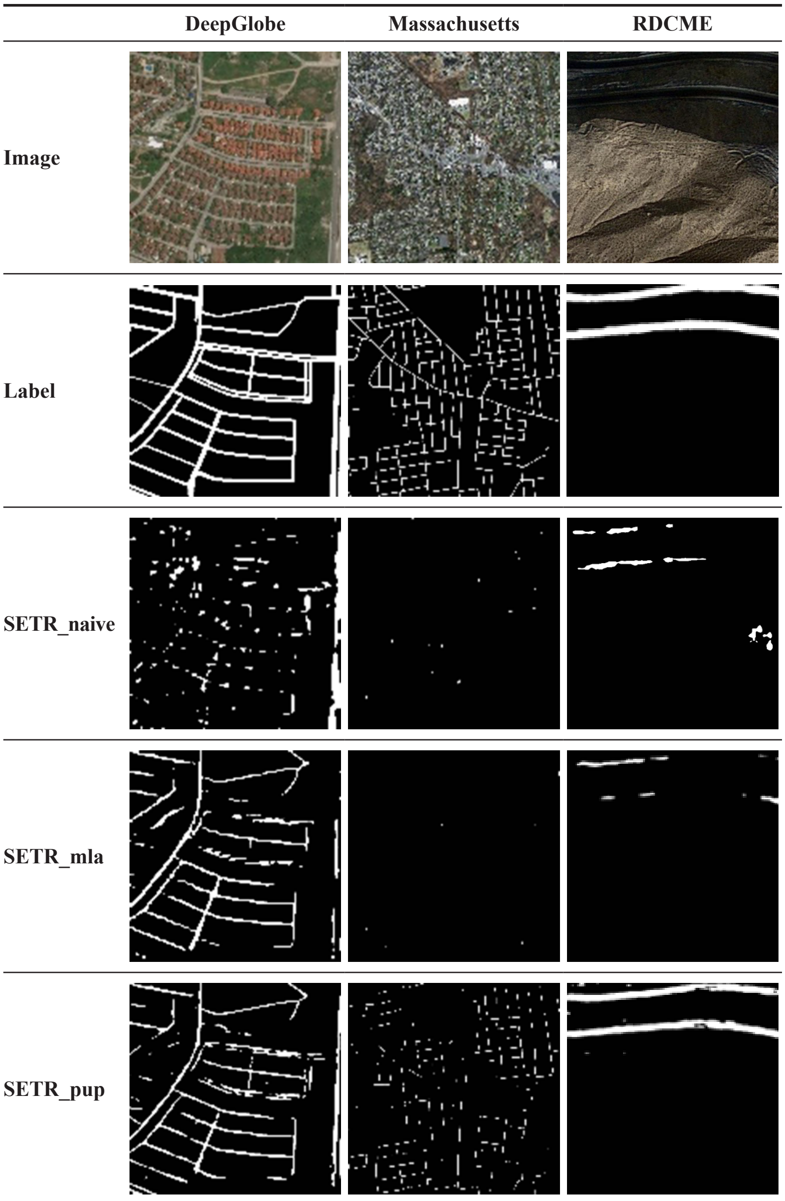
This study received funding from the following sources: the Open Research Fund of Key Laboratory of Smart Earth under KF2023YB02-07 and the Key Research and Development Program in Xianyang City under Grant L2023-ZDYF-QYCX-028.

Table 6. Evaluation metric values of four models for three datasets (%). Bolded data cells indicate the highest value for each evaluation metric within each dataset across all models

		Accuracy	Precision	Recall	F_1 score
FCN-8s	DeepGlobe	**97.11**	**87.14**	33.79	48.70
	Massachusetts	95.76	**83.26**	17.67	29.15
	RDCME	98.43	98.61	99.78	99.19
DPT	DeepGlobe	96.76	66.12	**41.37**	**50.89**
	Massachusetts	**96.31**	81.78	32.50	46.51
	RDCME	**98.94**	**99.07**	**99.84**	**99.46**
SegFormer	DeepGlobe	96.05	57.39	10.37	17.59
	Massachusetts	96.08	70.23	**35.79**	**47.42**
	RDCME	96.57	96.57	99.93	98.25
SETR_naive	DeepGlobe	95.78	35.33	4.69	8.28
	Massachusetts	95.05	40.51	0.20	0.41
	RDCME	93.75	98.35	95.13	96.71

DPT = dense prediction transformer; FCN = fully convolutional network; RDCME = Road Datasets in Complex Mountain Environments; SETR = SEgmentation TRansformer.

Table 5. Examples of road network extraction results for three datasets using different SETR variants.



SETR = SEgmentation TRansformer.

Table 7. Evaluation metric values of SETR variants for three datasets (%). Bolded data cells indicate the highest value for each evaluation metric within each dataset across all models

		Accuracy	Precision	Recall	F_1 score
SETR_naive	DeepGlobe	95.78	35.33	4.69	8.28
	Massachusetts	95.05	40.51	0.20	0.41
	RDCME	93.75	98.35	95.13	96.71
SETR_mla	DeepGlobe	96.44	**83.13**	15.32	25.87
	Massachusetts	95.08	65.83	1.03	2.02
	RDCME	97.26	97.29	99.96	98.60
SETR_pup	DeepGlobe	**96.55**	69.54	**26.65**	**38.53**
	Massachusetts	**95.65**	**80.63**	**15.68**	**26.26**
	RDCME	**98.23**	**98.52**	**99.67**	**99.09**

RDCME = Road Datasets in Complex Mountain Environments; SETR = SEgmentation TRansformer.

Disclosure statement

No conflict of interest was reported by the author(s).

References

Abdollahi, A., B. Pradhan, N. Shukla, S. Chakraborty and A. Alamri. 2020. Deep learning approaches applied to remote sensing datasets for road extraction: A state-of-the-art review. *Remote Sensing* 12(9):1444. https://doi.org/10.3390/rs12091444.

Alshehhi, R. and P. R. Marpu. 2017. Hierarchical graph-based segmentation for extracting road networks from high-resolution satellite images. *ISPRS Journal of Photogrammetry and Remote Sensing* 126:245–260. https://doi.org/10.1016/j.isprsjprs.2017.02.008.

Alshehhi, R., P. R. Marpu, W. L. Woon and M. D. Mura. 2017. Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks. *ISPRS Journal of Photogrammetry and Remote Sensing* 130:139–149. https://doi.org/10.1016/j.isprsjprs.2017.05.002.

Bae, Y., W.-H. Lee, Y. Choi, Y. W. Jeon and J. B. Ra. 2015. Automatic road extraction from remote sensing images based on a normalized second derivative map. *IEEE Geoscience and Remote Sensing Letters* 12(9):1858–1862. https://doi.org/10.1109/lgrs.2015.2431268.

Bagloee, S. A., M. Tavana, M. Asadi and T. Oliver. 2016. Autonomous vehicles: challenges, opportunities, and future implications for transportation policies. *Journal of Modern Transportation* 24(4):284–303. https://doi.org/10.1007/s40534-016-0117-3.

Bastani, F., S. He, S. Abbar, M. Alizadeh, H. Balakrishnan, S. Chawla, S. Madden and D. DeWitt. 2018. Roadtracer: Automatic extraction of road networks from aerial images, *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 18–22 June 2018, Salt Lake City, Utah (IEEE: Piscataway, New Jersey), pp. 4720–4728.

Buslaev, A. P., S. Seferbekov, V. Iglovikov and A. A. Shvets. 2018. Fully convolutional network for automatic road extraction from satellite imagery, *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops,* 18–22 June 2018, Salt Lake City, Utah. https://doi.org/10.1109/cvprw.2018.00035.

Cai, Y., H. He, K. Yang, S. Narges Fatholahi, L. Ma, L. Xu and J. Li. 2021. A comparative study of deep learning approaches to rooftop detection in aerial images. Canadian *Journal of Remote Sensing* 47(3):413–431. https://doi.org/10.1080/07038992.2021.1915756.

Chaudhuri, D., N. K. Kushwaha and A. Samal. 2012. Semi-automated road detection from high resolution satellite images by directional morphological enhancement and segmentation techniques. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 5(5):1538–1544. https://doi.org/10.1109/jstars.2012.2199085.

Chen, D., Y. Zhong, Z. Zheng, A. Ma and X. Lu. 2021. Urban road mapping based on an end-to-end road vectorization mapping network framework. *ISPRS Journal of Photogrammetry and Remote Sensing* 178:345–365. https://doi.org/10.1016/j.isprsjprs.2021.05.016.

Chen, Z., C. Wang, J. Li, B. Zhong, J. Du and W. Fan. 2021. Combined improved Dirichlet models and deep learning models for road extraction from remote sensing images. *Canadian Journal of Remote Sensing 47*(3):465–484. https://doi.org/10.1080/07038992.2021.1937087.

Chen, Z., L. Deng, Y. Luo, D. Li, J. Marcato Junior, J. Li, A. Nurunnabi, J. Li, C. Wang and D. Li. 2022. Road extraction in remote sensing data: A survey. *International Journal of Applied Earth Observation and Geoinformation* 112:102833–102833. https://doi.org/10.1016/j.jag.2022.102833.

Courtrai, L. and S. Lefèvre. 2016. Morphological path filtering at the region scale for efficient and robust road network extraction from satellite imagery. *Pattern Recognition Letters* 83:195–204. https://doi.org/10.1016/j.patrec.2016.05.014.

Demir, I., K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia and R. Raskar. 2018. DeepGlobe 2018: A challenge to parse the earth through satellite images, *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops,* 18–22 June 2018, Salt Lake City, Utah. https://ieeexplore.ieee.org/document/8575485.

Everingham, M., L. Van Gool, C.K.I. Williams, J. Winn and A. Zisserman. 2009. The Pascal Visual Object Classes (VOC) challenge. *International Journal of Computer Vision* 88(2):303–338. https://doi.org/10.1007/s11263-009-0275-4.

Goodfellow, I., Y. Bengio and A. Courville. 2016. *Deep Learning.* MIT Press: Cambridge Massachusetts, 800 p.

Goodfellow I. J., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio. 2014. Generative adversarial nets. Paper presented at the NIPS'14, Cambridge, MA, USA.

Grinias, I., C. Panagiotakis and G. Tziritas. 2016. MRF-based segmentation and unsupervised classification for building and road detection in peri-urban areas of high-resolution satellite images. *ISPRS Journal of Photogrammetry and Remote Sensing* 122:145–166. https://doi.org/10.1016/j.isprsjprs.2016.10.010.

He, H., D. Yang, S. Wang, Y. Zheng and S. Wang. 2019. Light encoder–decoder network for road extraction of remote sensing images. *Journal of Applied Remote Sensing* 13(3):1. https://doi.org/10.1117/1.jrs.13.034510.

Jain, J., J. Li, M. T. Chiu, A. Hassani, N. Orlov and H. Shi. 2023. Oneformer: One transformer to rule universal image segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 17–24 June 2023, Vancouver, Canada (IEEE: Piscataway, New Jersey), pp. 2989–2998.

Kestur, R., S. Farooq, R. Abdal, E. Mehraj, O. Narasipura and M. Mudigere. 2018. UFCN: A fully convolutional neural network for road extraction in RGB imagery acquired by remote sensing from an unmanned aerial vehicle. *Journal of Applied Remote Sensing* 12(1):1. https://doi.org/10.1117/1.jrs.12.016020.

Krylov, V. A. and J. D. B. Nelson. 2014. Stochastic Extraction of Elongated Curvilinear Structures With Applications. *IEEE Transactions on Image Processing* 23(12):5360–5373. https://doi.org/10.1109/tip.2014.2363612.

Kumbasar, D & Seker, D. 2023. Comparative Analysis of Different CNN Models for Building Segmentation from Satellite and UAV Images. Photogrammetric Engineering and Remote Sensing. 89. 97-105. 10.14358/PERS.22-00084R2.

Leninisha, S. and K. Vani. 2015. Water flow based geometric active deformable model for road network. *ISPRS Journal of Photogrammetry and Remote Sensing* 102:140–147. https://doi.org/10.1016/j.isprsjprs.2015.01.013.

Li, M., A. Stein, W. Bijker and Q. Zhan. 2016. Region-based urban road extraction from VHR satellite images using Binary Partition Tree. *International Journal of Applied Earth Observation and Geoinformation* 44:217–225. https://doi.org/10.1016/j.jag.2015.09.005.

Long, J., E. Shelhamer and T. Darrell. 2015. Fully convolutional networks for semantic segmentation, 2*015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 7–12 June 2015, Boston, Massachusetts (IEEE: Piscataway, New Jersey), pp. 3431–3440).

Ma T., H. Tan, T. Li, Y. Wu and Q. Liu. 2020. Road extraction method from GF-1 remote sensing images based on dilated convolution residual network with multi-scale feature fusion. *Laser & Optoelectronics Progress* 58:0228001.

Marchand, Y., L. Caraffa, R. Sulzer, E. Cledat and B. Vallet. 2023. Evaluating surface mesh reconstruction using real data. *Photogrammetric Engineering & Remote Sensing* 89(10):625–638. doi:10.14358/PERS.23-00007R3.

Mnih, V. 2013. *Machine Learning for Aerial Image Labeling.* Ph.D. thesis, University of Toronto, Toronto, Ontario, Canada, 103 pp. https://www.proquest.com/openview/215c1c8c5424a6facf8885f4ac00c575/1?pq-origsite=gscholar&cbl=18750 (last date accessed: 17/02/2024.

Oussama, M & Fatiha, M & Boukerch, I. 2023. Automatic Satellite Images Orthorectification Using K–Means Based Cascaded Meta-Heuristic Algorithm. Photogrammetric Engineering & Remote Sensing. 89. 30-39. 10.14358/PERS.22-00113R2.

Pan, D., M. Zhang and B. Zhang. 2021. A generic FCN-based approach for the road-network extraction from VHR remote sensing images—using OpenStreetMap as benchmarks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 14:2662–2673.

Poullis, C. and S. You. 2010. Delineation and geometric modeling of road networks. *ISPRS Journal of Photogrammetry and Remote Sensing* 65(2):165–181. https://doi.org/10.1016/j.isprsjprs.2009.10.004.

Ranftl, R., A. Bochkovskiy and V. Koltun. 2021. Vision transformers for dense prediction, *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 7–12 June 2015, Montreal, Canada (IEEE: Piscataway, New Jersey). https://doi.org/10.1109/ICCV48922.2021.01196.

Saito, S., and Y. Aoki. 2015. Building and road detection from large aerial imagery. *SPIE Proceedings Volume 9405, Image Processing: Machine Vision Applications VIII,* 8–12 February 2015, San Francisco, California (IS&T/SPIE). https://doi.org/10.1117/12.2083273.

Senthilnath, J., N. Varia, A. Dokania, G. Anand and J. A. Benediktsson. 2020. Deep TEC: Deep transfer learning with ensemble classifier for road extraction from UAV imagery. *Remote Sensing* 12(2):245. https://doi.org/10.3390/rs12020245.

Shamsolmoali, P., M. Zareapoor, H. Zhou, R. Wang and J. Yang. 2021. Road segmentation for remote sensing images using adversarial spatial pyramid networks. *IEEE Transactions on Geoscience and Remote Sensing* 59(6):4673–4688. https://doi.org/10.1109/tgrs.2020.3016086.

Subedi, M. R., C. Portillo-Quintero, S. S. Kahl, N. E. McIntyre, R. D. Cox and G. Perry. 2023. Leveraging NAIP imagery for accurate large-area land use/land cover mapping: A case study in central Texas. *Photogrammetric Engineering & Remote Sensing* 89(9):547–560. https://doi.org/10.14358/PERS.22-00123R2.

Taha, A. A. and A. Hanbury. 2015. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. *BMC Medical Imaging* 15(29). https://doi.org/10.1186/s12880-015-0068-x.

Tan, X., Z. Xiao, Q. Wan and W. Shao. 2021. Scale sensitive neural network for road segmentation in high-resolution remote sensing images. *IEEE Geoscience and Remote Sensing Letters* 18(3):533–537. https://doi.org/10.1109/lgrs.2020.2976551.

Tan, Y. Q., S. H. Gao, X. Y. Li, M. M. Cheng and B. Ren. 2020. Vecroad: Point-based iterative graph exploration for road graphs extraction, *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* 13–19 June 2020, Seattle, Washington (IEEE: Piscataway, New Jersey), pp. 8910–8918.

Valero, S., J. Chanussot, J. A. Benediktsson, H. Talbot and B. Waske. 2010. Advanced directional mathematical morphology for the detection of the road network in very high resolution remote sensing images. *Pattern Recognition Letters* 31(10):1120–1127. https://doi.org/10.1016/j.patrec.2009.12.018.

Van Etten, A., D. Lindenbaum and T. M. Bacastow. 2018. Spacenet: A remote sensing dataset and challenge series. arXiv:1807.01232 [cs.CV]. https://doi.org/10.48550/arXiv.1807.01232.

Wang, S., X. Mu, D. Yang, H. He and P. Zhao. 2021. Road extraction from remote sensing images using the inner convolution integrated encoder-decoder network and directional conditional random fields. *Remote Sensing* 13(3):465–465. https://doi.org/10.3390/rs13030465.

Wang, W., N. Yang, Y. Zhang, F. Wang, T. Cao and P. Eklund. 2016. A review of road extraction from remote sensing images. *Journal of Traffic and Transactions Engineering (English Edition)* 3(3):271–282.

Wang, Z., J. Q. Zheng, Y. Zhang, G. Cui and L. Li. 2024. Mamba-UNet: UNet-like pure visual mamba for medical image segmentation. arXiv:2402.05079 [eess.IV]. https://doi.org/10.48550/arXiv.2402.05079.

Wegner, J. D., J. A. Montoya-Zegarra and K. Schindler. 2015. Road networks as collections of minimum cost paths. *ISPRS Journal of Photogrammetry and Remote Sensing* 108:128–137.

Wu, J., R. Fu, H. Fang, Y. Zhang and Y. Xu. 2023. MedSegDiff-V2: Diffusion based medical image segmentation with transformer. arXiv:2301.11798 [eess.IV]. https://doi.org/10.48550/arXiv.2301.11798.

Wu, Q., F. Luo, P. Wu, B. Wang, H. Yang and Y. Wu. 2021 Automatic road extraction from high-resolution remote sensing images using a method based on densely connected spatial feature-enhanced pyramid. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 14:3–17. https://doi.org/10.1109/JSTARS.2020.3042816.

Xie, E., W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez and P. Luo. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems* 34:12077–12090.

Xu, H., H. He, Y. Zhang, L. Ma and J. Li. 2023. A comparative study of loss functions for road segmentation in remotely sensed road datasets. *International Journal of Applied Earth Observation and Geoinformation* 116:103159–103159. https://doi.org/10.1016/j.jag.2022.103159.

Xu, Z., Z. Shen, Y. Li, L. Xia, H. Wang, S. Li, S. Jiao and Y. Lei. 2021. Road extraction in mountainous regions from high-resolution images based on DSDNet and terrain optimization. *Remote Sensing* 13(1):90. https://doi.org/10.3390/rs13010090.

Yang, F., H. Wang, and Z. Jin. 2020. A fusion network for road detection via spatial propagation and spatial transformation. *Pattern Recognition* 100:107141. https://doi.org/10.1016/j.patcog.2019.107141.

Ye, Z., Y. Xu, H. Chen, J. Zhu, X. Tong and U. Stilla. 2020. Area-based dense image matching with subpixel accuracy for remote sensing applications: Practical analysis and comparative study. *Remote Sensing* 12(4):696. https://doi.org/10.3390/rs12040696.

Yu, Y., J. Wang, H. Guan, S. Jin, Y. Zhang, C. Yu, Pond, S. Xiao and J. Li. 2021. CS-CapsFPN: A context-augmentation and self-attention capsule feature pyramid network for road network extraction from remote sensing imagery. *Canadian Journal of Remote Sensing 47*(3):499–517. https://doi.org/10.1080/07038992.2021.1929884.

Yuan, J., D. Wang, B. Wu, L. Yan and R. Li. 2011. LEGION-based automatic road extraction from satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing* 49(11):4528–4538. https://doi.org/10.1109/tgrs.2011.2146785.

Zang, Y., C. Wang, Y. Yu, L. Luo, K. Yang and J. Li. 2017. Joint enhancing filtering for road network extraction. *IEEE Transactions on Geoscience and Remote Sensing* 55(3):1511–1525. https://doi.org/10.1109/tgrs.2016.2626378.

Zhang, X., W. Ma, C. Li, J. Wu, X. Tang and L. Jiao. 2020. Fully convolutional network-based ensemble method for road extraction from aerial images. *IEEE Geoscience and Remote Sensing Letters* 17(10):1777–1781. https://doi.org/10.1109/lgrs.2019.2953523.

Zhang, X., Jiang, Y., Wang, L., Han, W., Feng, R., Fan, R., & Wang, S. 2022. Complex mountain road extraction in high-resolution remote sensing images via a light roadformer and a new benchmark. *Remote Sensing*, *14*(19), 4729.

Zhang, Y., Z. Xiong, Y. Zang, C. Wang, J. Li and X. Li. 2019. Topology-aware road network extraction via multi-supervised generative adversarial networks. *Remote Sensing* 11(9):1017. https://doi.org/10.3390/rs11091017.

Zhang, Z., Q. Liu and Y. Wang. 2018. Road extraction by deep residual U-Net. *IEEE Geoscience and Remote Sensing Letters* 15(5):749–753. https://doi.org/10.1109/LGRS.2018.2802944.

Zhang, Z., C. Miao, C. Liu and Q. Tian. 2022. DCS-TransUperNet: Road segmentation network based on CSwin transformer with dual resolution. *Applied Sciences* 12(7):3511–3511. https://doi.org/10.3390/app12073511.

Zheng, S., J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P.H.S. Torr and L. Zhang. 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* 19–25 June 2021, virtual (IEEE: Piscataway, New Jersey), pp. 6881–6890).

Zhou, M., H. Sui, S. Chen, J. Wang and X. Chen. 2020. BT-RoadNet: A boundary and topologically-aware neural network for road extraction from high-resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing* 168:288–306.

In-Press Articles

SAT2BUILDING: LoD-2 Building Reconstruction from Satellite Imagery Using Spatial Embeddings
Philipp Schuegraf, Shengxi Gui, Rongjun Qin, Friedrich Fraundorfer, and Ksenia Bittner

The Aboveground Carbon Stock of Moso Bamboo Forests Is Significantly Reduced by Pantana phyllostachysae Chao Stress: Evidence from Multi-source Remote Sensing Imagery
Yuanyao Yang, Zhanghua Xu, Lingyan Chen, Wanling Shen, Haitao Li, Chaofei Zhang, Lei Sun, Xiaoyu Guo, and Fengying Guan

Cost-Effective High-Definition Building Mapping: Box-Supervised Rooftop Delineation Using High-Resolution Remote Sensing Imagery
Hongjie He, Linlin Xu, Michael A. Chapman, Lingfei Ma, and Jonathan Li

PROSAIL Modeling Coupled with Environmental Stress: Remote Sensing Retrieval of Multiple Dry Matters in the Canopy of Moso Bamboo Forests under the Stress of Pantana phyllostachysae Chao
Zhanghua Xu, Lei Sun, Yiwei Zhang, Huafeng Zhang, Hongbin Zhang, Fengying Guan, Haitao Li, Yuanyao Yang, and Chaofei Zhang

Wave Period and Direction Inversion from Marine X-band Radar Images using Spatiotemporal Feature Joint Learning
Li Wang, Hui Mei, Na Yang, Caiyun She, and Jian Qu