# Exploring The Application of Deep Learning to Road Detection in Remote Sensing Images

Kun Wang *

School of Mechanics and Civil Engineering, China University of Mining and Technology, Xuzhou, 221116, China

* Corresponding Author Email: 02220388@cumt.edu.cn

**Abstract.** Road information is a fundamental dataset in urban planning, traffic management, and related fields. However, challenges such as complex backgrounds, occlusions, and texture similarities with other linear features reduce the accuracy of road detection using remote sensing technology. This paper explores the advantages and recent developments in deep learning for enhancing road detection from remotely sensed images. The findings indicate that encoder-decoder semantic segmentation networks enable accurate road extraction; multi-scale feature fusion techniques enhance performance in occluded scenes; the integration of multi-task topology-constrained post-processing ensures road network connectivity and suppresses noise; and the adoption of weak supervision and lightweight models reduces labeling costs and improves deployment efficiency. Collectively, these advances will help remote sensing-based road detection reach new heights in terms of accuracy, intelligence, and breadth of application, driving the development of digital cities and intelligent transportation. This paper summarizes relevant techniques that support intelligent interpretation of remote sensing imagery and offers a valuable reference for future research in this field.

**Keywords:** deep learning, remote sensing images, road detection techniques.

## 1. Introduction

Road information is a fundamental dataset in various fields, including modern urban planning, traffic management, and Geographic Information System (GIS) updates. The automatic detection and extraction of roads from remote sensing images hold significant practical value [1]. With the increasing availability of high-resolution aerial and satellite imagery, there is ample data to support road detection efforts. However, challenges such as complex backgrounds, occlusions from trees and buildings, shadow interference, and texture similarities with other linear features (e.g., rivers, railways, etc.) significantly hinder the accurate extraction of roads from remote sensing images [1, 2].

With the development of artificial intelligence, deep learning has made significant breakthroughs in computer vision and remote sensing image analysis, thanks to its powerful feature learning and representation capabilities. These advances have provided new solutions for road detection in remote sensing imagery. Deep convolutional neural networks (CNNs) can automatically learn multi-level features from labeled data, overcoming the limitations of manually engineered features and demonstrating excellent performance in object detection and image segmentation tasks [3]. The adoption of deep learning methods has led to a substantial improvement in the accuracy of road detection from remote sensing images. For example, Yang et al. leveraged semantic segmentation network architectures—such as Fully Convolutional Networks (FCNs) and U-Net—to achieve end-to-end segmentation of road pixels [4]. Further improvements in detection performance and robustness under complex conditions have been achieved through the integration of multi-scale feature fusion, attention mechanisms, dilated convolutions, spatial pyramid pooling, and modules like Squeeze-and-Excitation (SE). Generative Adversarial Networks (GANs) have also been applied to enhance feature learning and segmentation quality. He et al. incorporated a multi-scale dilated convolution and pooling module into an encoder-decoder network, improving the F1 score on the Massachusetts Roads dataset by 2.6 percentage points [5]. Lin et al. introduced a channel attention mechanism to optimize feature representations, resulting in significant improvements in road continuity and extraction accuracy [6]. Through continued refinement, state-of-the-art methods now

achieve over 90% accuracy and F1 scores on public datasets [7, 2], enabling the reliable extraction of most road networks.

This paper presents a comprehensive literature review on road detection in remote sensing images using deep learning. It begins by outlining the overall development and key challenges associated with applying deep learning techniques to road extraction. The review then focuses on the evolution and application of core deep learning technologies—such as the U-Net architecture, attention mechanisms, dilated convolutions, Squeeze-and-Excitation (SE) modules, and multi-scale feature fusion—in the context of remote sensing-based road detection. These methods are categorized and analyzed in terms of their design principles, performance, applicable scenarios, and technical advancements, with representative studies cited for illustration. Finally, the paper summarizes the major accomplishments and current limitations of existing research and explores future trends and potential directions for continued development in this field.

## 2. The Development History of Deep Learning Remote Sensing Image Road Detection Algorithms

Before 2015, automatic road extraction from remote sensing images primarily relied on handcrafted features and heuristic rules, which exhibited limited adaptability in complex environments and occluded scenes (Fig.1). A notable early contribution came from Wang et al., who first introduced deep CNN into a large-scale road tracking framework by proposing a "DNN + finite state machine (FSM)" neural-dynamic model [8]. In this approach, the DNN identified road patterns within a sliding window, while the FSM controlled the tracking direction, enabling efficient road network tracking in high-resolution aerial imagery. This work laid the groundwork for a hybrid methodology combining deep learning-based feature extraction with traditional topological tracking.

Between 2018 and 2019, the growing availability of computational resources and public datasets—such as the Massachusetts Roads dataset—facilitated the adoption of FCNS and encoder-decoder architectures like U-Net in the remote sensing community. Henry et al. [9] applied a fully convolutional neural network (FCNN) to road segmentation in synthetic aperture radar (SAR) imagery, incorporating spatial tolerance rules to compensate for FCN's limited sensitivity to narrow linear structures, thereby pioneering the use of deep semantic segmentation in non-optical imagery.

From 2019 to 2021, with the end-to-end deep learning framework firmly established, research emphasis shifted toward feature recalibration and context awareness. Key strategies during this period included the introduction of attention mechanisms, channel weighting, multi-scale feature fusion, and residual dense connections. For instance, Ren et al. proposed DA-CapsUNet, which integrated capsule networks with dual attention mechanisms across both channel and spatial dimensions, enabling responsive detection in high-resolution settings and achieving an F1 score of 0.9504—the highest reported at the time [7]. Fan Jinhong's A-Res-U-Net incorporated attention gates to suppress redundant background features, outperforming Res-U-Net in completeness on the Massachusetts dataset [10]. Lin et al.'s Nested SE-Deeplab combined Squeeze-and-Excitation (SE) modules with multi-scale upsampling fusion, improving F1 and IoU scores by 2.4 and 2.0 percentage points, respectively [6]. Xu Miao et al. developed L-UNet, embedding SE modules and dilated convolutions into a lightweight backbone. The model was optimized for cloud-occluded scenes while reducing parameters to just one-fifth of those in D-LinkNet [11]. Wu et al.'s Dense-Global-Residual Net utilized dense connections and global spatial pyramid pooling to aggregate multi-scale contextual information, significantly enhancing road connectivity in occluded areas [3]. Jiang Na's improved ResNet-ASPP model—featuring ResNet101 with Atrous Spatial Pyramid Pooling (ASPP)—enhanced the receptive field and used sub-pixel convolutions to suppress upsampling artifacts, effectively restoring local road discontinuities [1].

In recent years, as models have continued to pursue higher accuracy, challenges related to annotation cost and computational overhead have come to the forefront. Two emerging research directions have addressed these issues: model lightweight and weak/semi-supervised learning. For

example, L-UNet and LCAN achieve mobile deployment-level inference speeds by incorporating MobileNet inverted residual blocks, depthwise separable convolutions, and dual attention mechanisms, all while maintaining leading IoU scores in cloud-occluded and complex scenes [11, 12]. ScRoadExtractor demonstrates the potential of weak supervision by generating pseudo-labels from sparse centerline scribble annotations using label propagation and boundary priors. This dual-branch network achieved IoU scores over 20% higher than traditional annotation-based methods, validating the feasibility of "high accuracy with minimal annotation" [13]. Additionally, Huang et al. employed GANS to generate synthetic road imagery under extreme lighting conditions, enhancing road detection in vehicle imagery [14].

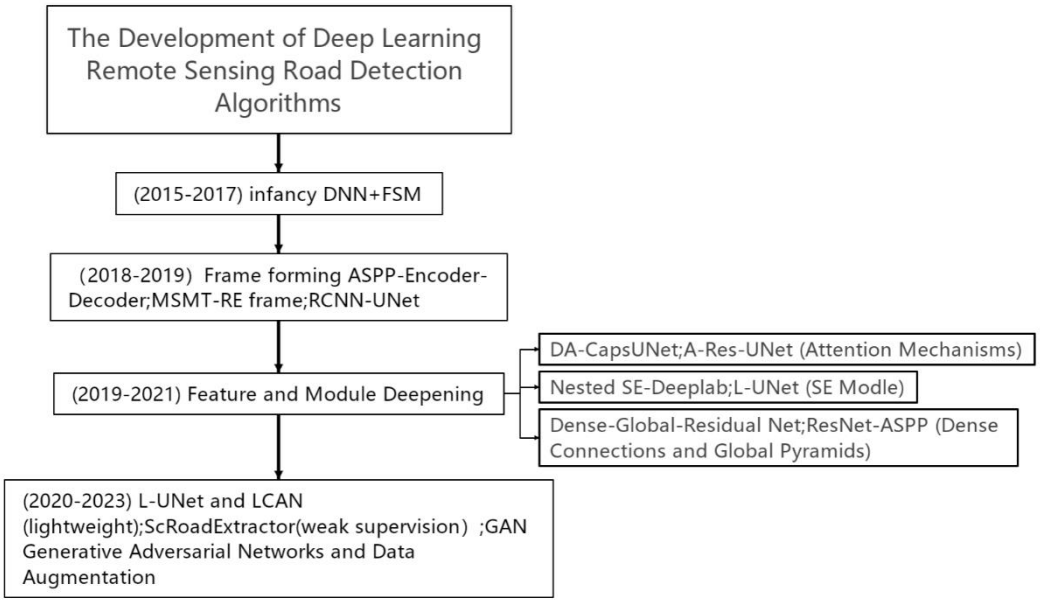


**Fig. 1** Development history of deep learning remote sensing image road detection algorithms

## 3. Research Progress on Deep Learning Remote Sensing Road Detection Methods

Traditional road extraction methods heavily rely on manually designed features and prior knowledge, resulting in insufficient utilization of high-dimensional features, complex algorithms, and limited adaptability to diverse scenarios. Deep learning methods, with their robust automatic feature learning capabilities, have significantly improved the accuracy and stability of road extraction.

Therefore, this section primarily focuses on the most representative deep learning methods in the field of remote sensing road detection in recent years, conducting an in-depth analysis of their key technical pathways. Four major categories of methods are analyzed in detail: first, encoder-decoder networks centered on U-Net and their extended structures, exploring their ability to restore road spatial details; second, context enhancement strategies based on dilated convolutions and spatial pyramid pooling to address the issue of capturing long-range information; third, various forms of multi-scale feature fusion methods, to enhance adaptability to changes in road width and different levels of semantic information; fourth, models incorporating attention mechanisms and SE modules to strengthen responses to target regions and channel feature selection.

### 3.1. U-Net architecture and encoder-decoder network

In remote sensing-based road detection, encoder-decoder semantic segmentation networks are widely adopted, with U-Net being one of the most influential architectures. It is symmetric downsampling–upsampling structure and cross-layer skip connections enable the preservation of both

global semantic information and local spatial details, making it a foundational model for many road extraction methods.

Roads in remote sensing imagery often exhibit characteristics such as narrow width and blurred edges, which pose challenges for traditional methods that frequently result in discontinuities or missed detections—particularly under low resolution or occlusion. U-Net addresses these issues by using skip connections to fuse shallow features capturing edge details with deep semantic features reconstructed by the decoder, thereby improving road edge localization and the continuity of extracted road networks. Building on this foundation, Lu et al. developed a multi-task, multi-scale road extraction framework based on U-Net, which performs both road region segmentation and centerline extraction. Both tasks share a common U-Net encoder to better preserve spatial structure [15]. Similarly, Yang et al. proposed a Recurrent Convolutional Neural Network U-Net (RCNN-UNet), which integrates recurrent convolutional modules into the U-Net architecture to model spatial context dependencies more effectively, enhancing robustness in noisy or occluded environments [4].

To address the limitations of the traditional U-Net—such as large parameter counts and insufficient depth in feature representation—residual networks (ResNets) have been integrated into its architecture. For example, Fan Jinhong et al. proposed the Res-U-Net model, which incorporates residual units into U-Net to enhance training efficiency and improve segmentation accuracy [10].

In addition, to better capture fine-grained road details, many approaches adopt pre-trained backbone networks—such as ResNet and DenseNet—as encoders, paired with U-Net-style decoders. This allows models to leverage the powerful representational capabilities of established architectures. A representative example is D-LinkNet, a U-shaped network that uses ResNet-34 as the encoder and connects it to the decoder through long skip connections. D-LinkNet has achieved excellent performance in remote sensing road extraction competitions [16].

Overall, the introduction of the U-Net architecture has marked the transition of remote sensing road detection into the era of end-to-end deep learning. Enhancements such as residual connections and pre-trained encoders have significantly improved extraction accuracy. However, relying solely on base encoder-decoder networks is often insufficient to address persistent challenges—such as road discontinuities and omissions in complex environments. This underscores the need for incorporating additional modules to strengthen contextual perception and structural reasoning. Table 1 summarizes the technical characteristics, limitations, and performance outcomes of U-Net-based and encoder-decoder network variants used in remote sensing road detection.

**Table 1.** Technical summary of the U-Net architecture and encoder-decoder network series framework

| Algorithm | Technical features | Defects | Results |
|---|---|---|---|
| **ASPP-Encoder-Decode(improved U-Net)** | First to use SSIM as loss suppression blurred boundaries. | ASPP increases the amount of calculations | Massachusetts Roads:F1 = 83.5%,SSIM = 0.893,F1 improved by 2.6% compared to the original U-Net. |
| **RCNN-UNet** | Inserting a convolutional unit into the U-Net decoder to capture long-range dependencies. | RNN units increase training difficulty; may still break on extremely long roads. | Surpassing nine baseline methods in two benchmark sets, Centerline F1 and Pixel F1 both ranked first. |
| **DA-CapsUNet** | The capsule network retains directional information and superimposes dual attention (channel + space) to extract roads end-to-end. | Capsule layer calculations are costly; narrow roads are still prone to missed inspections. | Large-scale dataset: Precision 0.9523 / Recall 0.9486 / F1 0.9504, comprehensively superior to 8 SOTA methods. |
| **ScRoadExtractor** | Using weak annotations of graffiti center lines, design a label propagation. | Depending on transmission quality; weak supervision performance is still slightly lower than full supervision. | Three global datasets: IoU outperforms classic scribble supervision by 20% and the latest weak supervision methods by at least 4%. |

### 3.2. Hollow convolution and spatial pyramid pooling

Dilated convolution (also known as hole convolution) is an effective technique for expanding the receptive field of road detection models while preserving spatial resolution, making it especially suitable for extracting long, continuous road features in remote sensing imagery. Unlike traditional downsampling-based approaches, dilated convolutions insert "holes" (i.e., spacing) between kernel elements, allowing the model to capture long-range contextual information without losing fine-grained details. Jiang Na's experiments demonstrate that, compared to standard convolutions, dilated convolutions significantly improve the model's ability to recognize elongated roads in complex backgrounds [1]. A notable extension of this concept is the Atrous Spatial Pyramid Pooling (ASPP) module, which captures multi-scale contextual information by applying dilated convolutions with varying dilation rates. When combined with pixel-level precision from decoder modules, ASPP enhances detail representation and continuity. For instance, incorporating ASPP into a U-Net-like structure resulted in an F1-score of 83.5% on the Massachusetts Roads dataset—a 2.6% improvement over the original U-Net—producing clearer and more complete segmentation results [5]. This highlights the value of multi-scale representations in overcoming occlusions and extracting continuous road networks.

However, excessive use of dilated convolutions can lead to the "gridding effect," where the receptive field becomes sparse and misses certain features. Therefore, it is essential to balance dilated

convolutions with standard convolutional operations or use hybrid structures. In addition to ASPP, spatial pyramid pooling (SPP) is another multi-scale strategy that pools features at multiple spatial scales to capture global and local contexts. Originally proposed for building extraction, SPP has also shown excellent results when applied to road detection. Networks combining encoder-decoder architectures with pyramid pooling demonstrated significantly higher accuracy on the Massachusetts Roads and Buildings dataset compared to standard U-Net models, confirming the generalizability of this approach [17]. Both ASPP and SPP are valuable for maintaining road continuity and repairing disconnections—particularly in complex environments or sparsely distributed road areas—by reducing both false positives and false negatives. Table 2 summarizes the technical characteristics, limitations, and application outcomes of models using dilated convolutions and spatial pyramid pooling, including performance comparisons across representative algorithms.

**Table 2.** Technical summary of the series of frameworks for hollow convolution and spatial pyramid pooling

| Algorithm | Technical features | Defects | Results |
|---|---|---|---|
| **Nested SE-Deeplab** | Based on Deeplab v3; decoder with multi-scale upsampling fusion + SE channel attention; core remains ASPP hollow convolution to obtain multi-scale features. | Large network size and high training memory requirements; still rely on dense labeling | Massachusetts: F1 ↑ 2.4 pct, IoU ↑ 2.0 pct vs FC-DenseNet; significant improvement in road integrity |
| **Dense-Global-Residual Net** | Construct a Dense + Global Spatial Pyramid Pooling (based on ASPP) module, with dense connections to retain multi-scale context; residual backbone to reduce information loss. | Dense connections and ASPP superposition, complex model | GF-2 and Massachusetts outperform baselines such as DeepLab v3+, U-Net, and D-LinkNet across the board, with outstanding connectivity in occlusion scenarios. |
| **ResNet-ASPP** | Using ResNet-101 as the backbone, multi-scale hollow convolution (ASPP) is used to extract ground object relationships; sub-pixel upsampling is combined to suppress interpolation noise. | Training requires a large amount of data to enhance anti-overfitting; ASPP has high computational overhead. | Experiments have verified that it "significantly improves road extraction accuracy and completeness," and is particularly effective for repairing locally broken sections. |
| **Topology-Thinning Dilated Framework** | Design "expansion module (void convolution) + information module" to form multi-scale features that expand the receptive field; combination loss maintains topological connectivity. | Still requires post-processing to eliminate noise; limited improvement on extremely narrow roads. | DeepGlobe: Accuracy ↑ 1.7–6.3 pct; Massachusetts: Accuracy ↑ 1.8–14.2 pct vs. comparison method |

### 3.3. Multi-scale feature fusion strategy

The width and morphology of roads in remote sensing imagery vary significantly—from broad highways to narrow, winding paths—posing a challenge for models that rely on single-scale feature

representations. Such models often struggle to detect roads of different widths simultaneously. As a result, multi-scale feature fusion has become a critical technique for enhancing road detection performance. This approach integrates features extracted at various resolutions and semantic levels, enabling the model to capture both fine-grained local details and larger-scale global structures.

While the skip connections in U-Net provide a basic form of multi-scale fusion by linking encoder and decoder features, researchers have developed more sophisticated strategies to further improve performance. For example, Lin et al. introduced a multi-scale upsampling fusion mechanism into the DeepLabv3 architecture. In their approach, feature maps from each encoder stage are upsampled to a common scale and then fused sequentially, preserving both low-level detail and high-level semantic information [6]. Similarly, in Lu et al.'s multi-task framework, multi-scale feature sharing is achieved by using a shared U-Net encoder for both road region segmentation and centerline extraction. The decoder then branches out into two task-specific heads. This design facilitates effective multi-scale learning and achieves superior performance compared to other deep learning-based methods—both quantitatively and qualitatively—on benchmark datasets [15]. These studies demonstrate that linking road features across different scales and structural forms for joint learning significantly enhances a model's ability to capture the topological continuity of road networks, particularly in complex or heterogeneous environments.

In addition to feature fusion at the network level, some studies have focused on result-level fusion and optimization. For instance, Wei et al. proposed a two-stage approach: an initial road probability map is first generated using a FCN, followed by iterative refinement using a simplified shallow FCN to enhance structural continuity [18]. This multi-stage, multi-scale processing pipeline significantly improved the completeness and coherence of road networks in large-scale remote sensing imagery. Compared to other methods, their approach achieved a 7% increase in the Connectivity metric and a 40% improvement in centerline extraction completeness [18].

Multi-scale feature fusion has now become a standard component in remote sensing-based road detection. At the network level, encoder-decoder architectures combined with techniques like dilated convolution and spatial pyramid pooling are employed to extract multi-scale contextual information. At the resulting level, strategies such as centerline–region fusion and morphological optimization further refine outputs to address the limitations of single-scale information. These techniques collectively enhance both the completeness and accuracy of road extraction. By integrating features across scales, models can more reliably detect roads of varying widths—from broad highways to narrow alleyways—and connect fragmented segments into a continuous road network. This continuity is essential for constructing accurate and connected road maps and is a critical metric in evaluating the quality of road extraction. Table 3 presents a technical summary of multi-scale feature fusion strategies, detailing the key characteristics, limitations, and application outcomes of representative algorithms.

**Table 3.** Technical summary of the multi-scale feature fusion strategy series framework

| Algorithm | Technical features | Defects | Results |
|---|---|---|---|
| **MSMT-RE: Multi-Scale & Multi-Task Deep Learning Framework** | Based on U-Net; integrating features of different resolutions in a multi-scale feature integration module, and designing a dual-branch architecture to simultaneously predict road regions. | Many branches and a large number of parameters, require high memory and training time. | In two publicly available road datasets (including Google Earth large maps), it outperformed all comparison models in both quantitative and qualitative terms, and led in connectivity metrics. |
| **Nested SE-Deeplab** | Based on Deeplab v3, multi-scale upsampling fusion is introduced, and SE channel attention is added to features at each scale to strengthen the complementary information between deep and shallow layers. | Large network volume and high memory requirements; still relies on pixel-level annotation | Massachusetts Roads: F1 ↑ 2.4%, IoU ↑ 2.0% compared to FC-DenseNet; road integrity significantly improved. |
| **Topology-Thinning Dilated Framework** | Design "expansion module (void convolution) + information module" to form multi-scale features that expand the receptive field; combination loss maintains topological connectivity. | Still requires post-processing to completely eliminate noise; limited improvement on extremely narrow roads. | DeepGlobe: Accuracy ↑ 1.7–6.3 pct; Massachusetts: Accuracy ↑ 1.8–14.2 pct vs. comparison method |

### 3.4. Introduction of Attention Mechanisms

In complex remote sensing imagery, improving road detection accuracy hinges on the model's ability to focus on relevant features that truly represent roads, while ignoring distracting or irrelevant background information.

Attention mechanisms play a critical role in addressing this challenge, particularly in scenarios where roads exhibit low contrast with their surroundings—such as areas shaded by trees or obscured by buildings. These modules guide the network to concentrate on residual road cues, thereby enhancing the continuity and reliability of road detection [12, 2]. Ren et al. proposed the Dual Attention Capsule U-Net (DA-CapsUNet), which integrates capsule networks with attention mechanisms for road region extraction [7]. The attention modules enable the model to generate more discriminative feature representations, improving its focus on relevant structures. On large-scale datasets, DA-CapsUNet achieved exceptional results—reporting a precision of 0.9523, recall of 0.9486, and an F1-score of 0.9504—surpassing eight other deep learning-based methods [7]. Zhou et al. tackled the specific issue of low contrast between asphalt roads and background buildings in high-

resolution images by introducing a Segmentation-based Deep Separable Graph Convolutional Network (SGCN) [2]. Their method first uses depthwise separable convolutions to extract both channel and spatial features and then applies graph convolutions to capture global dependencies within feature maps. To enhance edge awareness, the Sobel operator is used to extract edge information, which is then employed to construct the adjacency matrix for the graph—effectively functioning as an edge-guided attention mechanism that reinforces subtle road boundary cues [2]. These examples illustrate that, whether through explicit attention modules or hybrid architectures incorporating graph structures and capsules, the underlying goal remains the same: enabling the model to intelligently filter and emphasize useful information, thereby allowing it to "focus" on road features amid complex and noisy backgrounds.

Another widely adopted approach is the joint attention mechanism, which simultaneously applies attention to both channel and spatial dimensions. For example, Kang Liangfang et al. proposed a lightweight convolutional attention network that integrates spatial and channel attention modules within a convolutional neural network [12]. In this architecture, spatial attention determines where important features are located, while channel attention identifies which types of features are most informative. The combination of these two attention types enables the model to suppress irrelevant semantic information, enhance the selection of road-relevant features, and reduce the impact of low-level noise. By selectively amplifying critical information and filtering distractions, this dual-attention approach improves the model's ability to detect roads in complex scenes. Given its effectiveness, it is foreseeable that attention mechanisms will become increasingly integral to various aspects of remote sensing image analysis and continue to evolve alongside road extraction techniques. Table 4 provides a technical summary of attention mechanisms, outlining their core characteristics, limitations, and application outcomes across different algorithms.

**Table 4.** Technical summary of the series of frameworks introducing attention mechanisms

| Algorithm | Technical features | Defects | Results |
|---|---|---|---|
| **L-UNet** | Mobile Inverted Bottleneck + Depthwise Separable Conv; combining SE channel attention and dilated convolutions, designed for cloud occlusion scenarios; model lightweighting. | In terms of high-quality cloud-free imagery, accuracy is slightly inferior to heavy networks; the generalizability of synthetic cloud datasets remains to be verified. | DeepGlobe extended set IoU ↑ 1.97 pct; real cloud scene IoU ↑ 19.47 % compared to D-LinkNet, ↑ 31.87 % compared to U-Net; parameters only 1/5 of D-LinkNet |
| **A-Res-U-Net** | Integrating Attention Gate (spatial attention) into ResNet-U-Net, dynamically filtering irrelevant features, improving training efficiency and segmentation accuracy | The model is still biased; improvements to fully occluded scenes are limited. | Massachusetts Roads Overall accuracy and connectivity are higher than the original U-Net and Res-U-Net, and training time is reduced. |
| **LCAN:Lightweight Convolution Attention Network** | Simultaneously introducing spatial + channel joint attention in a deep separable convolution framework. | The paper does not provide uniform metrics; further evaluation is needed for complex occlusion extreme scenarios. | The authors report significant improvements over models without attention mechanisms. |

### 3.5. SE Module (Squeeze-and-Excitation) and Channel Attention

The Squeeze-and-Excitation (SE) module is a channel-focused attention mechanism that enhances feature representation by explicitly modeling inter-channel dependencies. In remote sensing road detection, the SE module has been widely adopted due to its simple design and proven effectiveness, contributing significantly to improvements in extraction performance.

The SE module operates in two main stages: squeeze, which applies global average pooling to each feature channel to capture global contextual information; and excitation, which uses fully connected layers and non-linear activations to generate a set of channel-wise weights. These weights are then used to recalibrate the feature maps by enhancing channels relevant to road features while suppressing those unrelated or prone to noise. This selective emphasis improves the model's focus on road-relevant structures. Lin et al. integrated the SE module into the DeepLab v3 framework, developing a nested SE-DeepLab model for road extraction from high-resolution remote sensing imagery [6]. By embedding SE attention into features at different convolutional layers, the model dynamically adjusted the importance of each channel, thereby enhancing the continuity and completeness of extracted road networks. Furthermore, as shown in Jiang Na's work [1], SE principles have also been extended to fuse multi-source features. A residual-style squeeze-and-excitation network was used to extract and combine features from both spatial and spectral dimensions, demonstrating the module's versatility across different types of input data.

Empirical results confirm that adding SE modules can reduce fragmented false positives and improve road connectivity, leading to higher extraction completeness [6]. The SE module's lightweight nature and compatibility with various encoder-decoder architectures make it easy to integrate into existing deep learning pipelines. Given its effectiveness, the SE module is expected to remain a valuable component in the continued advancement of remote sensing-based road detection methods.

## 4. Challenges and potential solutions

Despite recent advancements, current research in remote sensing road detection still faces several key challenges. Models remain highly dependent on large volumes of annotated training data, yet pixel-level labeling is costly and time-consuming. Additionally, regional differences in imagery can hinder generalization across geographic areas [13, 2]. Roads that are heavily occluded by tree canopies or buildings, as well as narrow secondary roads, remain particularly difficult to detect [2]. Furthermore, models often lack robustness under varying material types and lighting conditions, while some of the most accurate architectures suffer from high computational complexity and large parameter counts, limiting their practical deployment in resource-constrained environments [11, 12]. To address these limitations, researchers have increasingly turned to weak supervision learning, lightweight network design, and multi-source data fusion as promising development directions.

Given the high cost of acquiring pixel-level annotations, several studies have explored weakly supervised approaches to reduce reliance on manual labeling. Notably, Wei and Ji proposed ScRoadExtractor, a weakly supervised method that uses only easily annotated thin-line centerlines as training supervision [13]. By developing a label propagation algorithm, their approach extends sparse centerline annotations into dense pseudo-labels for road regions and incorporates boundary priors to guide training. The method employs a dual-branch encoder-decoder network to process both the propagated labels and boundary features. Tested across three high-resolution road datasets from different geographic regions, ScRoadExtractor achieved results comparable to fully supervised methods [13]. These findings demonstrate that through carefully designed label propagation strategies and multi-task learning, weak supervision can approach the performance of traditional supervised methods. As such, the adoption of weak and semi-supervised learning techniques holds strong potential for significantly reducing annotation costs, thereby offering a scalable and effective solution to the data scarcity problem in high-precision road detection.

To enable the deployment of road detection models on resource-constrained platforms—such as mobile devices, embedded systems, or large-scale online services—model lightweight has become an increasingly important research focus. Xu et al. proposed L-UNet, which integrates MobileNet's inverted residual modules and depthwise separable convolutions to drastically reduce model complexity. The resulting network has only 20% of the parameters of the original large-scale model, yet achieves superior performance, particularly in cloud-obscured scenarios [11]. These efforts

support the practical implementation of road detection technology, providing a solid foundation for fast, large-scale updates of road data in real-world applications.

In parallel, remote sensing-based road detection is evolving to integrate multi-source data and multi-task learning to generate more comprehensive representations of traffic networks. For example, Shao et al. [19] applied an improved object detection framework—combining Faster R-CNN, feature pyramid networks (FPNs), and attention mechanisms—to detect road intersections from remote sensing imagery. These detected intersections are then matched to existing vector map intersections to perform automatic alignment between imagery and vector road network data. This approach not only supports rapid map updates, particularly for crowdsourced platforms but also demonstrates the practical value of road detection in cartographic and geospatial applications.

In summary, deep learning-enabled road detection from remote sensing imagery is rapidly advancing. By incorporating weak supervision, topological priors, novel network architectures, and engineering optimizations, researchers are steadily enhancing the accuracy, robustness, and applicability of road extraction techniques. These diverse innovations are expanding the scope of remote sensing analysis and driving the field toward more intelligent and integrated interpretations of geospatial data.

## 5. Conclusion

Research on road detection in remote sensing images using deep learning has made significant progress in recent years, with numerous studies demonstrating that deep learning methods have a decisive advantage over traditional methods. The main contributions of existing research are as follows: first, end-to-end road segmentation models have been developed that can automatically learn road features from complex remote sensing images and achieve high-precision pixel-level extraction; second, multi-scale and attention modules have been introduced to overcome the challenges posed by the narrow and easily obstructed nature of road targets, thereby enhancing the completeness and continuity of extraction results; third, multi-task, topological constraints, and post-processing techniques have been employed to ensure the connectivity of road networks, significantly reducing discontinuities and noise; fourth, gradually focusing on weak supervision and lightweight approaches, making valuable explorations to reduce data annotation costs and enhance model practicality. It can be said that the continuous evolution of deep learning technology is leading remote sensing road detection from early crude and fragmented extraction results toward a new era of precise and reliable automated extraction.

Looking ahead, deep learning holds considerable potential for further advancement and broader application in the field of remote sensing-based road detection. As larger and more diverse remote sensing datasets are constructed and shared, models will be better positioned to learn more robust and discriminative feature representations. The adoption of weakly supervised learning will reduce reliance on fully annotated data, enabling a shift toward a "big data with weak labels" training paradigm that balances scalability with efficiency. Emerging architectures such as GNN and Transformer-based self-attention mechanisms offer promising solutions for modeling long-range spatial dependencies, which is essential for generating globally coherent road networks. Additionally, integrating road detection with related tasks—such as change detection, map updating, and autonomous driving—can help establish a comprehensive traffic perception system that fuses multi-source, multi-level data from satellite to ground perspectives. As early as 2016, scholars identified “deep learning feature representation” and “weakly supervised object detection” as two pivotal directions for the evolution of remote sensing target detection. These predictions have since been partially realized and are expected to continue driving progress in the field. With the ongoing maturation of deep learning theory and remote sensing technology, road detection from remote sensing imagery is poised to achieve new levels of accuracy, intelligence, and real-world impact. These advancements will play a critical role in enabling the development of smart cities, intelligent transportation systems, and automated geospatial analytics.

## References

[1] Jiang Na. Research on Road Extraction Methods for Remote Sensing Images Based on Deep Learning [D]. Lanzhou Jiaotong University, 2023. DOI: 10.27205/d.cnki.gltec.2023.000009.

[2] Zhou, G., Song, W., Zheng, Z., Wang, S., & Yang, J. (2022). Split Depth-Wise Separable Graph-Convolution Network for Road Extraction in Complex Environments from High-Resolution Remote-Sensing Images. IEEE Transactions on Geoscience and Remote Sensing, 60, 5528513.

[3] Wu, Q., Tong, X., Huang, X., Jin, X., & Deng, S. (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.

[4] Yang, X., Yao, X., Shi, Y., & Li, H. (2019). Road Detection and Centerline Extraction via Deep Recurrent Convolutional Neural Network U-Net. IEEE Transactions on Geoscience and Remote Sensing, 57(9), 7209-7220.

[5] He, H., Yang, F., & Zhang, X. (2019). Road Extraction by Using Atrous Spatial Pyramid Pooling Integrated Encoder-Decoder Network and Structural Similarity Loss. Remote Sensing, 11(9), 1015.

[6] Lin, Y., Di, K., Li, X., Cui, Y., Gao, B., & Yue, Z. (2020). Road Extraction from Very-High-Resolution Remote Sensing Images via a Nested SE-Deeplab Model. Remote Sensing, 12(18), 3015.

[7] Ren, Y., Yu, Y., & Guan, H. (2020). DA-CapsUNet: A Dual-Attention Capsule U-Net for Road Extraction from Remote Sensing Imagery. Remote Sensing, 12(18), 3042.

[8] Wang, J., Song, Y., Zhang, Y., & Li, J. (2015). Road network extraction: a neural-dynamic framework based on deep learning and a finite state machine. International Journal of Remote Sensing, 36(12), 3144-3169.

[9] Henry, C., Azimi, S. M., & Merkle, N. (2018). Road Segmentation in SAR Satellite Images with Deep Fully Convolutional Neural Networks. IEEE Geoscience and Remote Sensing Letters, 15(12), 1867-1871.

[10] Fan Jinhong. Research and Application of Road Recognition Technology Based on Remote Sensing Images [D]. Beijing University of Posts and Telecommunications, 2021. DOI: 10.26969/d.cnki.gbydu.2021.001013.

[11] Xu Miao, Li Yuanxhang, Zhong Juanjuan, et al. L-UNet: A Lightweight Cloud-Occluded Road Extraction Network [J]. Journal of Image and Graphics, 2021, 26(11): 2670-2679.

[12] Kang Liangfang. Research on Road Extraction from High-Resolution Remote Sensing Images Based on a Lightweight Convolutional Attention Network [D]. Guangxi Normal University, 2021. DOI: 10.27036/d.cnki.ggxsu.2021.000552.

[13] Wei, Y., & Ji, S. (2022). Scribble-Based Weakly Supervised Deep Learning for Road Surface Extraction from Remote Sensing Images. IEEE Transactions on Geoscience and Remote Sensing, 60, 5520413.

[14] Huang, X. Research on road detection based on deep learning [D]. Guangxi University, 2020. DOI: 10.27034/d.cnki.ggxiu.2020.000781.

[15] Lu, X., Feng, J., Song, Y., & Tao, Y. (2019). Multi-Scale and Multi-Task Deep Learning Framework for Automatic Road Extraction. IEEE Transactions on Geoscience and Remote Sensing, 57(11), 9362-9377.

[16] Peng Bo. Research on Deep Learning-Based Remote Sensing Image Road Information Extraction Algorithm [D]. University of Electronic Science and Technology of China, 2019.

[17] Liu, Y., Peng, J., Zhang, Y., & Liu, H. (2019). Automatic Building Extraction on High-Resolution Remote Sensing Imagery Using Deep Convolutional Encoder-Decoder with Spatial Pyramid Pooling. IEEE Access, 7, 128774-128786.

[18] Wei, Y., Zhang, K., & Ji, S. (2020). Simultaneous Road Surface and Centerline Extraction from Large-Scale Remote Sensing Images Using CNN-Based Segmentation and Tracing. IEEE Transactions on Geoscience and Remote Sensing, 58(12), 8919-8931.

[19] Shao Xin. Research on vector data and remote sensing image registration based on deep learning [D]. Wuhan University, 2021. DOI: 10.27379/d.cnki.gwhdu.2021.000047.