

TSNet: A Two-stage Segmentation Network Guided by Uncertainty for Remote Sensing Image Road Extraction

Jiajia Liu^{1,*}, Xuan Zhao², Wenxiang Dong², Jianhua Liu², Xiaoguang Tu²

¹Civil Aviation Flight University of China, Center for Faculty Development and Teaching Evaluation, Guanghan 618307, China

²Civil Aviation Flight University of China, College of Aviation Electronic and Electrical Engineering, Guanghan 618307, China

* Corresponding author

Abstract: Road extraction from high-resolution remote-sensing images plays a vital practical role in multiple application scenarios including urban layout planning and autonomous driving systems. Nevertheless, current road-extraction approaches still suffer from several bottlenecks under complex ground environments, including indistinct object boundaries, missing fine-grained features and fragmented road outputs. Aiming to mitigate the above defects, this paper puts forward an uncertainty-guided two-stage segmentation network termed TSNet. The proposed framework divides road segmentation into two successive subtasks: global-level semantic understanding and local boundary optimization. A cross-layer residual feature enhancement module (CRM) is adopted to extract and aggregate multi-scale feature information. On the basis of multi-scale fused features, the first segmentation stage produces preliminary probability maps, and an uncertainty-aware learning module (UALM) is introduced to measure the confidence degree for each pixel-wise prediction. In the subsequent second stage, the local refinement module (LRM) takes the generated uncertainty map as guidance to dynamically concentrate on regions prone to prediction errors, and implements adaptive optimization for road edge areas. Quantitative experiments on the DeepGlobe dataset prove that the presented model achieves competitive performance in maintaining road connectivity, yielding Accuracy, Precision, Recall, F1-score and Intersection over Union (IoU) of 98.68%, 83.72%, 82.36%, 82.85% and 71.07%, respectively. Further validation on the CHN6-CUG dataset obtains an Accuracy of 96.10%, Precision of 80.48%, Recall of 79.56%, F1-score of 79.37% and IoU of 66.20%. When compared with state-of-the-art competing algorithms, our method achieves optimal results for all five evaluation metrics. In addition, qualitative visualization results reveal that the model exhibits prominent strengths in retaining connected road topology and recovering intricate edge contours, which effectively enhances the robustness and prediction accuracy of road extraction for complex real-world scenes.

Keywords: Remote sensing imagery; Road extraction; Semantic segmentation; Two-stage network; Uncertainty-aware.

1. Introduction

The development of remote sensing technology has made the acquisition of high-resolution imagery increasingly convenient. Precisely extracting road networks from such imagery has become a key challenge in the field of remote sensing information processing. However, due to complex background interference, illumination variations, occlusions, and the intricate topological structure and multi-scale nature of roads in remote sensing images, traditional road extraction methods struggle to balance global structural coherence with local boundary accuracy[1].

The introduction of deep learning technology marks a paradigm shift in remote sensing image road extraction tasks. Early attempts were mainly based on traditional image processing and machine learning methods but with limited effectiveness[2]. U-Net[3], with its encoder-decoder structure and skip connections, was applied in biomedical image segmentation and quickly migrated to the remote sensing field. However, facing the characteristics of roads in remote sensing images, such as long distances, narrowness, and complex topology, the simple skip connections of U-Net suffer from a semantic gap when fusing high- and low-level features. Its fixed-size receptive field is difficult for effectively modeling linear targets with variable shapes and complex backgrounds like roads, often leading to discontinuous extraction and blurred boundaries[4]. To obtain a larger receptive field for capturing richer global context, DeepLabv3+[5] introduced the Atrous Spatial Pyramid Pooling (ASPP) module. This

module uses parallel convolutions with different dilation rates to capture multi-scale information, improving the model's robustness to large-area occlusions and complex scenes. However, the backbone network of DeepLabv3+ loses a significant amount of spatial detail in deep layers, resulting in insufficient ability to recover fine boundaries and slender roads. Its decoder part is relatively simple, making it difficult to precisely restore pixel-level localization from highly abstract features, easily causing artifacts or adhesion at fine structures like intersections. To address the linear and long-range dependency characteristics of roads, D-LinkNet[6] made improvements. It adds a central dilated convolution module between the encoder and decoder, which expands the receptive field through stacked dilated convolutions while maintaining high-resolution feature maps. However, D-LinkNet is essentially a single-stage, end-to-end prediction paradigm. This paradigm treats the model as a "black box," outputting the final result in one pass, lacking a mechanism for reflection and correction during the intermediate process. Models based on self-attention mechanisms like Swin Transformer[7] and SegFormer[8] are widely used. With their powerful global modeling capability, they can better understand the large-scale topology of roads. However, Transformer-based models typically require massive data for pre-training, have high computational resource demands, and their local detail perception capability is sometimes inferior to carefully designed convolutional networks. They may still be insufficient in pixel-level precise localization, prone to errors on small objects or thin edges. SEG-Road[9] explicitly

optimizes the continuity of road networks by introducing graph convolutional or connectivity structural constraints, alleviating discontinuity problems caused by occlusions. Such methods often rely on post-processing or complex graph reasoning, increasing computational complexity, and their generalization ability in complex urban scenes still has room for improvement. DSNet[10] explores dense fusion of multi-scale contexts by designing a Densely connected Spatial Pyramid Pooling module, while dual-branch networks like DDRNet[11] attempt to achieve a better balance between "detail" and "semantics" by maintaining a high-resolution branch to preserve spatial details while using another deep branch to extract semantic context. These works partially alleviate specific problems but often increase structural complexity and computational cost. Their core optimization focus remains on designing more powerful feature extractors and does not change the "one-time output" inference paradigm. MARENet[12] achieves leading performance on multiple public datasets through multi-scale attention and enhanced feature recalibration mechanisms. HDRoad [13] designs a hybrid attention mechanism in the encoder and introduces directional morphological priors in the decoder, aiming to reconstruct road geometric features more finely. However, whether it is early architectural improvements (DSNet, DDRNet) or recent global mechanisms and specialized optimizations (Swin, MARENet, HDRoad), these methods lack a mechanism for self-assessment during inference and dynamic, targeted optimization based on the assessment results. Models cannot quantify the uncertainty of their predictions at each pixel location, thus failing to adaptively focus limited computational resources on the most difficult and error-prone regions (e.g., areas with heavy occlusion, complex overpass shadow areas). This leads to difficulties in ensuring the topological connectivity of road networks while pursuing overall accuracy metrics. Misjudgments in high-uncertainty regions remain the main cause of road discontinuities and detail loss.

To overcome the inherent drawbacks of existing frameworks, namely one-off prediction, identical computation for all pixels and absence of internal self-evaluation, this study develops an Uncertainty-Guided Two-stage Segmentation Network (TSNet). Instead of adopting the conventional one-off end-to-end mapping strategy, TSNet splits the road segmentation task into two coordinated sub-stages: global semantic perception and local boundary refinement. Within this framework, an Uncertainty-Aware Learning Module (UALM) is deployed to automatically evaluate pixel-level prediction confidence, which further offers dynamic guidance for resource distribution inside the Local Refinement Module (LRM). Such design targets higher boundary precision, better retention of fine-scale details as well as sound topological connectivity for road extraction amid complicated scenarios, and offers an alternative solution to tackle prevailing defects in existing research. The major contributions of this paper are outlined as follows:

1. This work devises an uncertainty-driven two-stage segmentation architecture. Road segmentation is divided into coarse prediction and uncertainty-driven fine-grained segmentation, so as to realize adaptive handling for complicated scenarios. Following the processing logic of "global analysis prior to local adjustment; reliable regions prior to ambiguous regions", the proposed structure

overcomes the shortcomings of classic end-to-end models that apply identical computation without internal evaluation. It enables the network to concentrate on high-error-risk zones and carry out targeted optimization.

2. A Cross-layer Residual feature enhancement Module (CRM) is presented. This module mitigates spatial detail loss occurred in deep backbone features, and generates feature embeddings integrating high-level semantic knowledge and low-level geometric cues for follow-up segmentation procedures. Via explicit cross-layer feature aggregation and feature recalibration operations, the module can well retain road edge features and the structural information of narrow road segments.

3. This paper puts forward the Uncertainty-Aware Learning Module (UALM) together with the Local Refinement Module (LRM). The pixel-wise segmentation confidence is calculated based on the entropy derived from predicted probability distributions, generating interpretable dynamic attention weights to support local optimization. Driven by the uncertainty map, the network adaptively pays more attention to areas with high error probability, which substantially boosts boundary segmentation precision and the connectivity of extracted road networks.

4. Experiments carried out on the DeepGlobe dataset demonstrate the superior performance of our approach in extracting complete road structures. The corresponding Accuracy, Precision, Recall, F1-score and IoU are 98.68%, 83.72%, 82.36%, 82.85% and 71.07%. Evaluations on the CHN6-CUG dataset further yield the following metrics: Accuracy of 96.10%, Precision of 80.48%, Recall of 79.56%, F1-score of 79.37% and IoU of 66.20%.

2. Method

2.1. Overall Network Architecture

This study develops TSNet, an architecture adopting uncertainty-driven progressive refinement, which divides road segmentation into two dedicated processing phases. The first phase is dedicated to preliminary identification of global semantic structures, while the second phase implements adaptive optimization for local boundary details. Such design concept originates from thorough analysis on the drawbacks of existing segmentation networks. Conventional end-to-end architectures encounter great difficulty in striking a balance between globally consistent semantic representation and precise local boundaries when coping with complex real-world scenarios.

As illustrated in Figure 1, the overall TSNet architecture complies with the processing pipeline of "Encoding-Coarse Decoding-Uncertainty Estimation-Fine Decoding".

The network first performs preliminary feature extraction through the CRM module, converting the image into four feature maps of different dimensions:

$$F_1 \in \mathbb{R}^{128 \times H/4 \times W/4} \quad (1)$$

$$F_2 \in \mathbb{R}^{256 \times H/8 \times W/8} \quad (2)$$

$$F_3 \in \mathbb{R}^{512 \times H/8 \times W/8} \quad (3)$$

$$F_4 \in \mathbb{R}^{1536 \times H/16 \times W/16} \quad (4)$$

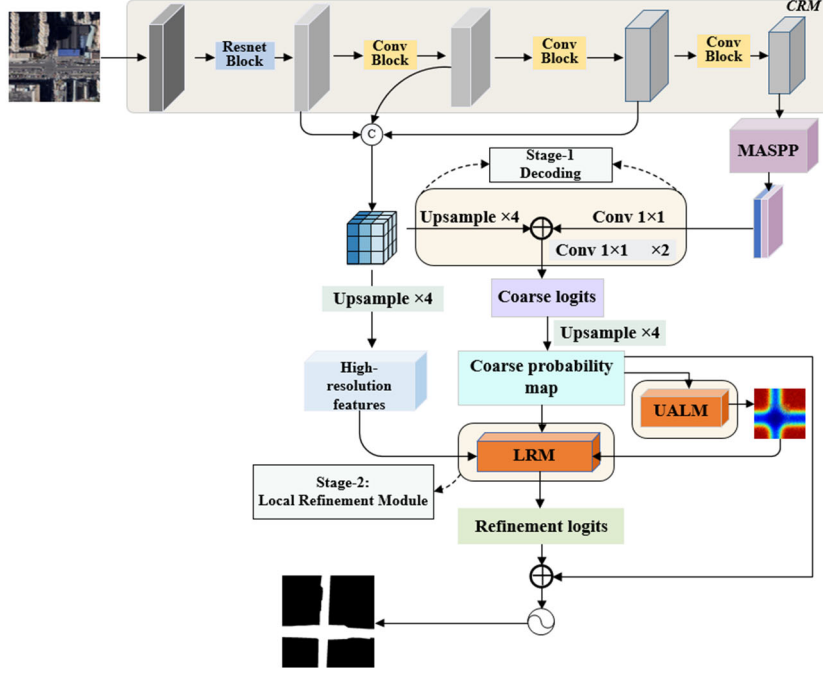


Figure 1. TSNet Network Architecture

where F_1 , F_2 , F_3 are rich in spatial detail information, and F_4 contains high-level semantic context.

Subsequently, the high-level semantic feature map F_4 is input into the modified Multi-scale Atrous Spatial Pyramid Pooling (MASPP) module. This unit makes use of dilated convolutions with multiple dilation rates. It can substantially raise the segmentation accuracy for semantic tasks, and meanwhile avoid semantic degradation as well as discontinuous segmentation predictions for road targets. Different from the original ASPP module, MASPP introduces dilated convolution whose dilation rate is set to 24 for further expanding the receptive field of the network.

To fully utilize multi-scale information, this paper adaptively fuses mid- and low-level features. F_2 and F_3 undergo upsampling to attain the spatial resolution consistent with F_1 before being concatenated across channel dimensions. The first-stage decoder combines the MASPP output with the upsampled high-level features to generate the preliminary segmentation prediction. The decoder adopts a lightweight design, containing two 3×3 convolutional blocks, each followed by batch normalization and ReLU activation. Through 4 upsampling steps and a sigmoid activation

function, the coarse segmentation probability map is obtained. The coarse segmentation probability map is processed by the UALM to obtain the uncertainty map. The high-resolution feature map, coarse segmentation probability map, and uncertainty map are convolved and passed through the LRM to obtain the refined feature map. The refined feature map is fused with the coarse segmentation probability map and activated to yield the final prediction map.

2.2. Cross-layer Residual feature enhancement Module

To address the loss of road spatial detail in deep semantic features due to consecutive downsampling in the encoder, this work proposes the Cross-layer Residual Feature Enhancement Module (CRM), Figure 2 presents the specific structure of this module. The core target of CRM lies in building strengthened feature expressions that fuse bottom-layer geometric characteristics and top-layer semantic knowledge during early feature extraction by means of clear cross-layer feature exchange and combination, offering better preliminary data for later coarse segmentation and local detail refinement procedures.

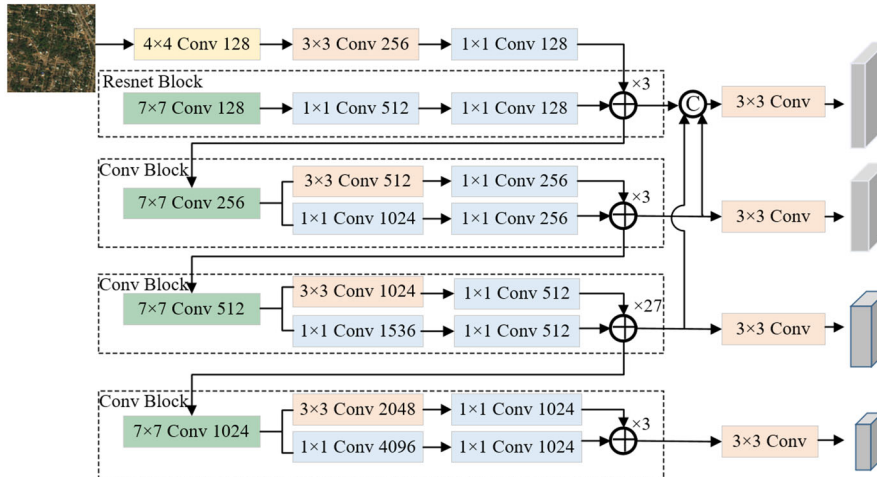


Figure 2. CRM Module

The core design concept of the CRM module lies in feature reuse and gradient direct flow. The module introduces residual connections and feature fusion mechanisms at multiple stages, performing element-wise addition or channel-wise recalibration of shallow spatial detail features and deep semantic features. Such operation supplements the structural information derived from low-level feature layers, and meanwhile preserves the discriminative capability of high-level features. The corresponding mathematical formulation is given as follows:

$$F_{out}=C(F_{high},R(F_{low}))\in R^{C\times H\times W} \quad (5)$$

where F_{high} and F_{low} represent high-level semantic features and low-level spatial features, respectively; $R(\cdot)$ denotes the residual connection and feature mapping function; $C(\cdot)$ is the cross-layer feature fusion operation.

Incorporating the CRM module into the network strengthens feature representation capacity, and meanwhile greatly optimizes the model's performance for capturing thin

and broken road structures within complex urban environments. Equipped with this module, the network can mitigate road missing issues triggered by over-abstract semantic features. Besides, it generates higher-grade feature representations to support the subsequent coarse segmentation and local refinement phases.

2.3. Uncertainty-Aware Learning Module

To resolve low prediction confidence occurring at road boundaries, shadow-occluded areas and other hard-to-process regions during the coarse segmentation phase within complicated scenarios, this study designs the Uncertainty-Aware Learning Module (UALM). Figure 3 demonstrates its overall processing pipeline. The core objective of UALM is to produce pixel-level interpretable uncertainty metrics through computing the information entropy derived from coarse-segmentation probability distributions. This resulting uncertainty map serves as an adaptive spatial attention weight, precisely guiding the subsequent Local Refinement Module to perform targeted enhancement on high-error-risk regions.

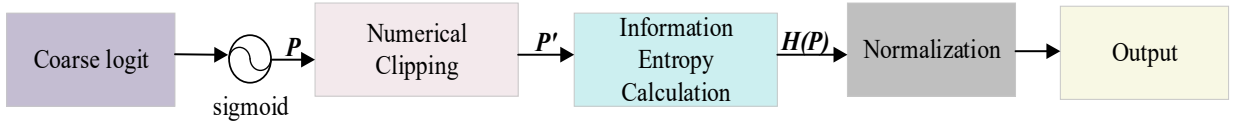


Figure 3. UALM Module

The core of the UALM module is to quantify uncertainty by calculating the information entropy of the predicted probability distribution. Given the coarse segmentation logits L_c output by the first-stage decoder, they are first converted into a probability map via the Sigmoid function:

$$P_c=\sigma(L_c)\in[0,1]^{H\times W} \quad (6)$$

For binary classification, the prediction uncertainty of pixel (i, j) can be quantified by its binary entropy:

$$H(p_{ij})=-p_{ij}\log_2 p_{ij}-(1-p_{ij})\log_2(1-p_{ij}) \quad (7)$$

where p_{ij} denotes the probability of pixel (i,j) being classified as a road pixel. This function has the following mathematical properties:

1. Monotonic symmetry: Monotonically increasing on the interval $[0, 0.5]$ and decreasing on $[0.5, 1.0]$.
2. Extremum properties: $H(0.5)=1$ (maximum uncertainty), $H(0)=H(1)=0$ (complete certainty).
3. Differentiability: Continuously differentiable on the interval $(0,1)$, facilitating gradient backpropagation.

To ensure numerical stability, probability values are clipped in practical implementation:

$$p'_{ij}=\text{clip}(p_{ij},\epsilon,1-\epsilon),\epsilon=10^{-8} \quad (8)$$

Finally, the normalized uncertainty map is obtained:

$$U_{ij}=\frac{H(p'_{ij})}{\log_2 2}\in[0,1] \quad (9)$$

The normalization factor $\log_2 2$ corresponds to the maximum entropy for binary classification, constraining the uncertainty map U_{ij} within the range $[0,1]$ for subsequent processing.

2.4. Local Refinement Modules

Serving as the key component within the second phase of TSNet, the LRM realizes the segmentation mechanism from coarse outputs toward fine-grained predictions. Relying on the coarse-segmentation probability map, high-resolution spatial features and uncertainty estimation map yielded by the first-stage computation, this module conducts adaptive optimization for vague zones, boundary details and regions with semantic fuzziness in prediction outputs. Following the “identify-focus-optimize” cognitive computing logic, LRM mitigates typical defects existing in conventional end-to-end segmentation networks for remote-sensing image analysis, including blurry boundaries, missing fine features and fractured small-target structures. The network layout of the LRM is presented in Figure 4.

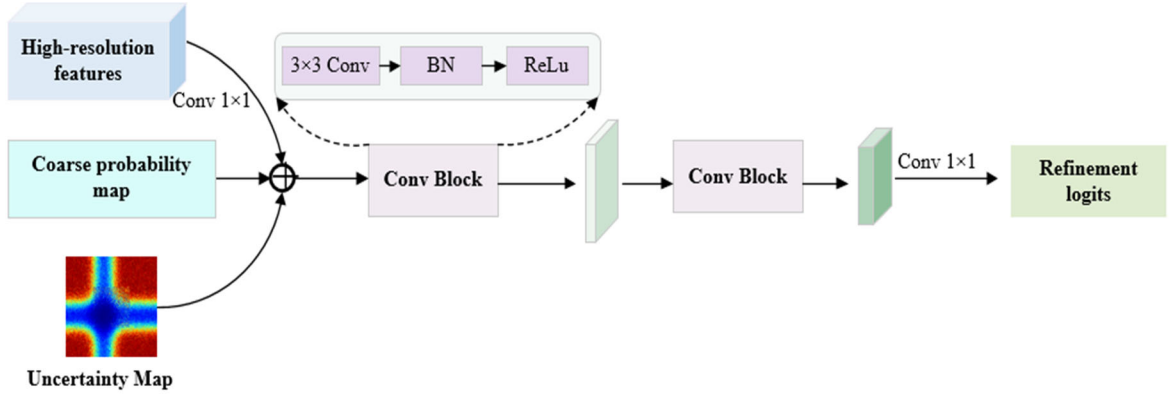


Figure 4. Network Architecture of the LRM Module

The LRM module starts with element-wise multiplication operations. Specifically, the dimension-reduced high-resolution feature F_d , which is obtained via up-sampling and channel compression on low-level backbone features, is multiplied with the coarse-segmentation probability map P_c as well as the uncertainty map U . This operation strengthens semantic guidance and enables the model to place emphasis on critical regions:

$$F_{fused} = F_d * P_c * U \quad (10)$$

By means of the uncertainty weight U , this operation adaptively imposes suppression on background areas. It enables the module to pay more attention to potential error-prone zones, namely high-uncertainty regions, and cuts down redundant calculations over areas with high prediction confidence. The coarse-segmentation probability map supplies global semantic priors to steer the refinement procedure, whereas the uncertainty map serves as dynamic attention weights to realize spatially non-uniform distribution of computing resources.

The fused feature F_{fused} is fed into two successive Conv Block units. Every block comprises a 3×3 convolution operator, a batch normalization layer and ReLU nonlinear activation function. These components extract local contextual cues and boost the representation capability of feature maps. Following the two convolution cycles, a 1×1 convolutional layer projects the feature into the target channel dimension (generally set to 1) and outputs the ultimate refined feature map $F_{refined}$:

$$F_{refined} = \text{Conv}_{1 \times 1}(F_{enhanced}) \in \mathbb{R}^{1 \times H \times W} \quad (11)$$

Guided by the uncertainty map, the LRM module dynamically tunes computational weights and allocates computing resources toward high-value zones, namely regions with high uncertainty and high error risk. It combines spatial detail information, semantic prior knowledge and confidence estimation outputs to promote the reliability and precision of local correction operations. Interacting with coarse segmentation outputs, LRM builds a closed-loop feedback mechanism. This design prevents gradient vanishing triggered by excessively long gradient propagation paths during end-to-end training, and meanwhile improves the model's adaptability to complex scenarios. Benefiting from its lightweight architecture, LRM maintains favorable computational efficiency. It realizes high-quality transformation from coarse segmentation outputs to

fine-grained segmentation results, without bringing substantial extra inference cost.

Leveraging multi-source feature collaboration, uncertainty-driven attention mechanisms and local enhancement sub-networks, the LRM conducts fine-grained correction for road segmentation outputs of remote-sensing images, and further elevates model reliability and prediction precision under complex ground conditions.

2.5. Loss Function

In this work, a hybrid loss function integrating Binary Cross-Entropy (BCE) loss[14] and Dice loss[15] is adopted to strengthen the segmentation capability of the proposed model. BCE loss can capture structural features for retaining texture information of images, whereas Dice loss pays more attention to boundary features within high-frequency areas. This combination enhances the learning capability for road target regions, improves the network's adaptability to imbalanced categories, and ensures the integrity of global features while extracting fine edge features.

In this study, a hybrid loss function combining Binary Cross-Entropy (BCE) loss[14] and Dice loss[15] is employed to boost the segmentation performance of the presented model. BCE loss is capable of capturing structural characteristics so as to preserve textural information from input images. By contrast, Dice loss places greater emphasis on boundary features distributed in high-frequency domains. Such joint loss formulation strengthens the network's learning capacity for road target areas, ameliorates its adaptability toward category imbalance, and guarantees intact global feature representation in the meantime of extracting delicate edge characteristics. The specific expression of this loss function is as follows:

$$L_{Dice} = 1 - \frac{2 \sum_i^{H \times W} \phi_{true}^i \phi_{pred}^i + 1}{\sum_i^{H \times W} (\phi_{true}^i)^2 + \sum_i^{H \times W} (\phi_{pred}^i)^2 + 1} \quad (12)$$

$$L_{BCE} = -\frac{1}{N} \sum_{i=1}^N \left(\gamma \phi_{true}^i \log(\phi_{pred}^i) + (1 - \phi_{true}^i) \log(1 - \phi_{pred}^i) \right) \quad (13)$$

$$L_{Hybrid} = L_{BCE}(\delta_{true}, \delta_{pred}) + L_{Dice}(\delta_{true}, \delta_{pred}) \quad (14)$$

where N is the number of pixels in the sample. The binary class label ϕ_{true}^i for pixel i represents the ground truth label of the sample, taking a value of 0 or 1. The expected

probability ϕ_{pred}^i for sample classification ranges between 0 and 1. Through the loss functions L_{Road} and L_{Edge} (targeting edge details and sample labels, respectively) and the prediction result loss functions δ_{true} and δ_{pred} (targeting overall road texture), the accuracy of the segmentation model and edge information extraction is improved to varying degrees.

3. Experiments

3.1. Evaluation Metrics

Researchers generally utilize five assessment criteria to quantify how well models extract road targets from remote sensing imagery: Precision, Recall, Intersection over Union, F1 metric and overall Accuracy[16]. Specifically, Precision quantifies the ratio of accurately inferred road pixels to all pixels labeled as roads by the model, and its computational expression is given below:

$$\text{Pre} = \frac{TP}{TP+FP} \quad (15)$$

Recall quantifies the ratio of accurately predicted road pixels to all real road pixels marked in ground-truth maps. Its calculation formula is:

$$\text{Recall} = \frac{TP}{TP+FN} \quad (16)$$

Accuracy represents the proportion of accurately classified pixels among all pixel samples. Its calculation formula is:

$$\text{Acc} = \frac{TP+TN}{TP+FN+FP+TN} \quad (17)$$

As a comprehensive evaluation index, F1 metric corresponds to the harmonic mean derived from Precision and Recall, with the specific calculation formula shown below:

$$F_1 = \frac{2 \times \text{Pre} \times \text{Recall}}{\text{Pre} + \text{Recall}} = \frac{2TP}{2TP+FN+FP} \quad (18)$$

As a core evaluation metric, IoU quantifies the overlap extent of predicted road domains and annotated ground truth regions, and its computational equation is defined as follows:

$$\text{IoU} = \frac{TP}{TP+TN+FP} \quad (19)$$

Higher scores of the above metrics correspond to superior road extraction performance.

In the road prediction results, TP (True Positive) and FP (False Positive) represent the total number of correctly identified road pixels and incorrectly identified road pixels, respectively. FN (False Negative) reflects the number of true positive pixels incorrectly classified as negative, while TN (True Negative) represents correctly identified background

pixels.

3.2. Experimental Datasets and Settings

The experimental environment settings fully consider the comparability of results within the same group. All experiments are conducted under identical environments and parameters to ensure fair comparison. Experiments are performed on a Windows 11 system using a single NVIDIA GeForce RTX 4070 (12GB GDDR6) GPU, with a Hygon C86-3G (OPN: 3350) CPU. The experimental environment employs Python 3.12 for programming, and deep neural networks are constructed relying on PyTorch 2.5.1 together with CUDA 12.1.

All experimental verifications in this study are performed on two public road extraction datasets, namely the DeepGlobe dataset and the CHN6-CUG road dataset. As a commonly used remote sensing road dataset, DeepGlobe covers diverse scene types and provides accurate road region annotations for each image. The images of this dataset possess a fixed resolution of 1024×1024 pixels and a ground sampling distance of 0.5 meters per pixel. For experimental implementation, the dataset is partitioned into 4980 samples for model training, 250 samples for validation, and 996 samples for model testing. Differently, the CHN6-CUG road dataset targets typical urban cluster scenarios in China. All images in this dataset are cropped to 512×512 pixels with a consistent spatial resolution of 0.5 meters per pixel, which is specially constructed for domestic urban road extraction tasks. This dataset is randomly split into 3608 training samples and 903 test samples to support network training and performance evaluation. To strengthen the network's anti-interference capability and optimize the generalization performance during the training process, multiple data augmentation strategies are adopted for input image preprocessing, including random color distortion, random rotation transformation, image translation, as well as horizontal and vertical flipping operations.

In this work, the Adam optimizer is employed for model training with a batch size of 4 and an initial learning rate of 2×10^{-4} . A multi-step learning rate decay strategy is implemented to facilitate efficient and stable training. Specifically, the learning rate will be reduced by half if the training loss fails to decline obviously within six successive epochs, and the training procedure will terminate automatically once the learning rate falls below 5×10^{-8} . Furthermore, to boost the road extraction accuracy and speed up network convergence, all backbone networks adopt pre-trained weights from the ImageNet-1K dataset for parameter initialization.

3.3. Ablation Study

To quantitatively evaluate the contribution of each core module in the proposed TSNet network to the overall network performance and verify the effectiveness of the "two-stage optimization" and "uncertainty guidance" mechanisms, a systematic ablation study is conducted on the CHN6-CUG road dataset. Detailed numerical results from the conducted experiments are illustrated in Table 1.

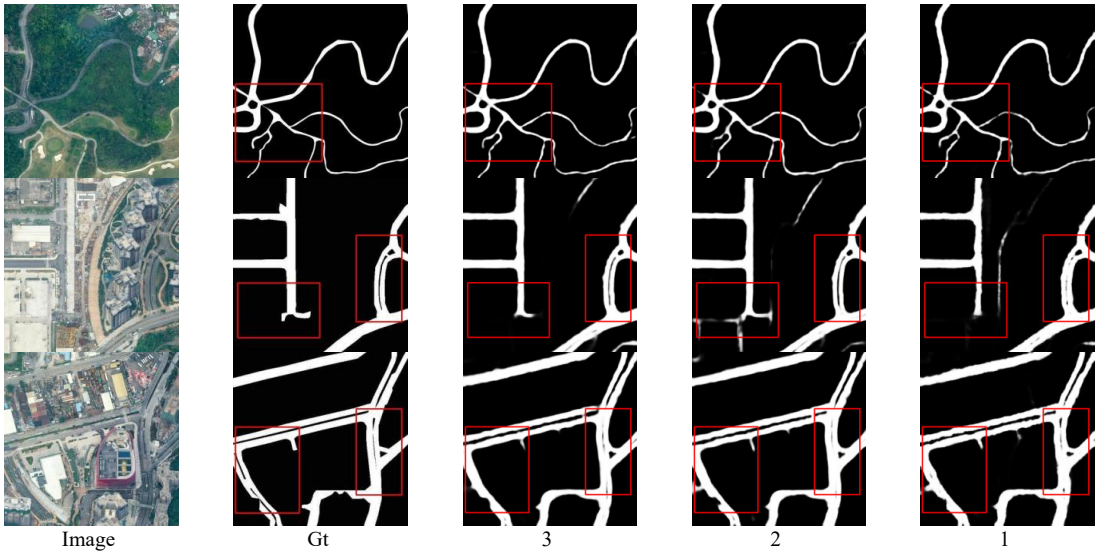
Table 1. Experimental ablation results of the proposed method on the CHN6-CUG dataset

Group	Model	Two-Stage	UALM	LRM	F1-score (%)	IoU (%)
1	UncertaintyNet	√	√	×	78.84	65.53
2	RefineNet	√	×	√	79.15	65.87
3	TSNet	√	√	√	79.37	66.20

From Table 1, it can be observed that after introducing the LRM but removing uncertainty guidance, the model's F1-score increases from 78.84% to 79.15%, and IoU increases from 65.53% to 65.87%. Such performance gain demonstrates that local refinement based on high-resolution features can effectively remedy detail loss generated in the coarse-segmentation phase, even in the absence of explicit uncertainty guidance. It preliminarily verifies that the refinement module itself delivers tangible benefits for boosting boundary precision and structural completeness. Comparing the results of Group 2 and Group 3, after embedding the uncertainty-guided mechanism back into the network — that is, adopting the uncertainty map to steer feature fusion inside the refinement module — the model obtains a higher F1-score of 79.37% and an IoU of 66.20%. These metrics are 0.22% and 0.33% higher than Group 2 correspondingly. This observation proves that the uncertainty

guidance module is not dispensable. By closely cooperating with the local refinement module, it enables the network to precisely target high-value zones, i.e., regions with high uncertainty. Rather than triggering indiscriminate computation over every pixel, this mechanism drives the refinement module to prioritize optimization for boundaries and discontinuous road segments where the model yields low-confidence predictions and tends to make mistakes. Consequently, competitive segmentation performance can be achieved under constrained computational budgets.

To qualitatively investigate how the core modules within TSNet affect segmentation outputs at a fine-grained level and confirm the practical effects brought by the “uncertainty guidance” and “local refinement” mechanisms under complex scenarios, visual comparison experiments are carried out based on the CHN6-CUG road dataset.

**Figure 5.** Visual Comparison Chart of the Ablation Study

As illustrated in Figure 5, the uncertainty estimation module substantially improves the model’s discrimination performance under interfering conditions. Meanwhile, the local refinement module driven by uncertainty further boosts the restoration capacity for geometric details. When these two components function jointly inside the TSNet framework, the model can obtain high precision without sacrificing recall metrics. This offers a feasible solution to tackle complex background disturbance and detail-missing defects existing in remote-sensing image-based road extraction tasks.

3.4. Comparative Experiments

To fully verify the performance advantages of the designed method in remote sensing road extraction tasks, comparative experiments against several advanced semantic segmentation algorithms are carried out in this work. Five prevalent quantitative metrics, including Accuracy, Precision, Recall, F1-score, and Intersection over Union (IoU), are employed to conduct a comprehensive evaluation of the model’s overall

performance.

For a fair and comprehensive performance assessment of the TSNet model, Table 2 summarizes its comparative scores on five key indicators based on the DeepGlobe dataset.

As illustrated in Table 2, the proposed TSNet outperforms other comparative methods and achieves state-of-the-art results on all core evaluation metrics. Specifically, TSNet achieves an accuracy of 98.68%, which is 0.19% higher than MARENet (98.49%), indicating higher reliability in overall pixel classification. On the Precision metric, TSNet reaches 83.72%, demonstrating that the model can effectively suppress background noise and significantly reduce false positives in road extraction. TSNet achieves a Recall of 82.36%, ensuring that the continuity of slender roads and complex intersections is not missed. The F1-score of the proposed model reaches 82.85%, and the IoU reaches 71.07%. TSNet improves upon the suboptimal MARENet by 2.16% and 2.85% in F1-score and IoU, respectively. These results indicate that through its unique two-stage processing

mechanism, uncertainty-aware learning, and local refinement module, TSNet can effectively address common issues in road extraction such as blurred boundaries, discontinuities, and

omissions, achieving an effective balance between precision and completeness.

Table 2. Comparative analysis of road extraction performance based on the DeepGlobe dataset

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	IoU (%)
A2FPN[17]	98.01	75.77	76.01	75.89	61.05
DeepLabv3+[5]	97.98	76.06	73.29	73.79	59.29
D-LinkNet[6]	98.16	77.73	76.48	76.78	62.62
DSNet[10]	98.08	76.28	75.62	75.59	61.17
BASNet[18]	98.34	80.03	78.49	79.01	65.63
FFNet[19]	98.04	75.92	75.68	75.79	60.80
UNet3+[20]	97.90	77.68	72.35	73.08	59.62
UNetformer[21]	97.97	73.51	78.14	75.75	60.95
MARENet[12]	98.49	81.39	80.77	80.69	68.22
improve DeepLab3+[22]	97.89	79.65	81.66	80.66	67.35
TSNet	98.68	83.72	82.36	82.85	71.07

To further verify the above quantitative analysis conclusions from a visual dimension, Figure 6 shows visual

comparison charts on the DeepGlobe road dataset.

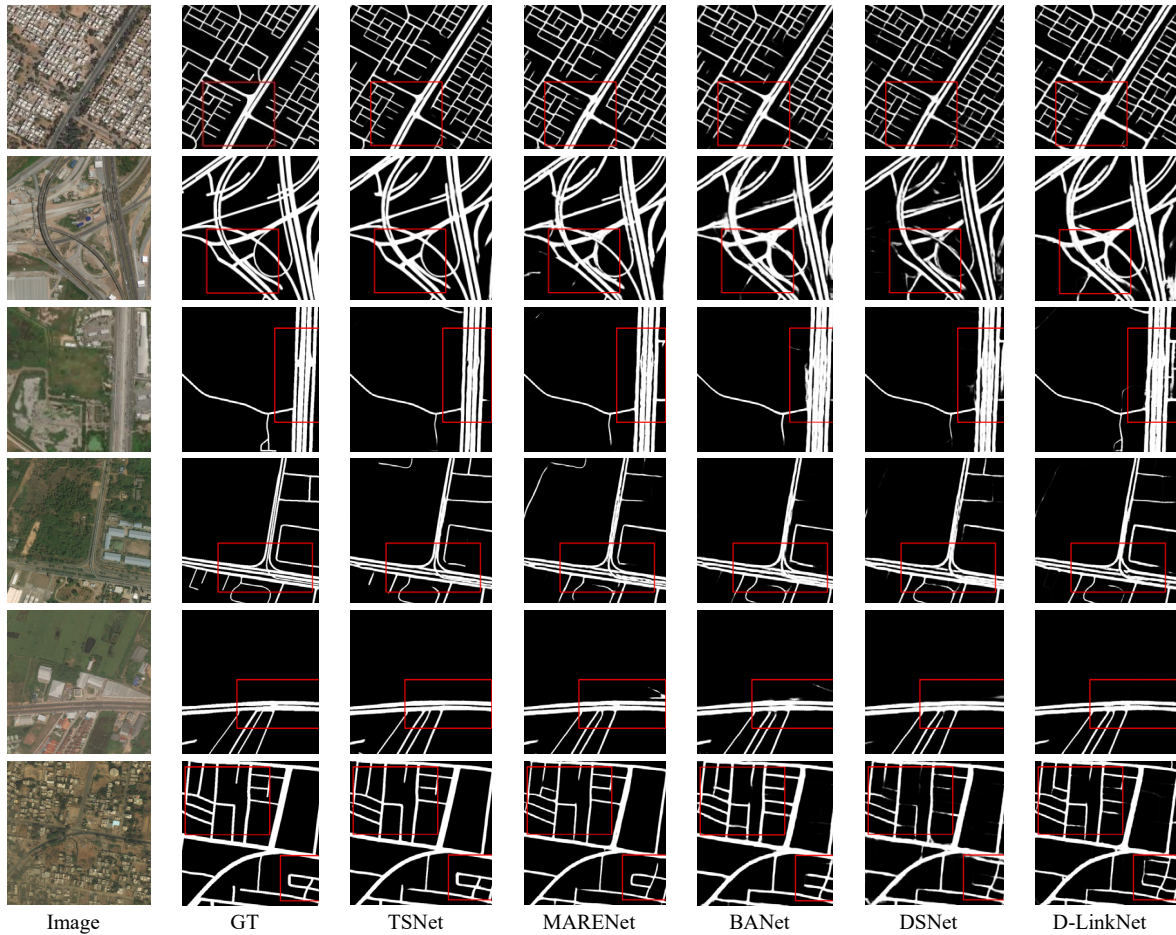


Figure 6. Visual assessment of diverse road extraction approaches on the DeepGlobe dataset

According to the qualitative visualization outcomes presented in Figure 6, TSNet shows superior adaptability and extraction accuracy when coping with diverse complex remote-sensing scenarios. It greatly reduces artifacts caused by fragmented road segments and prevents tiny branch roads from being missed, finally generating complete, smooth and structurally consistent road extraction outputs. Benefiting from its two-stage refinement workflow and uncertainty-driven strategy, TSNet can reconstruct occluded

road contours more satisfactorily within shadow-covered and low-contrast regions, which further strengthens the topological connectivity of extracted road networks. Roads extracted by TSNet deliver smoother and more regular edge outlines. The model effectively mitigates jagged edges and burr-like artifacts that frequently emerge in the prediction results of other competing approaches. Such performance advancement mainly stems from high-precision boundary feature reconstruction implemented by the local refinement

module.

Further comparative experiments are implemented on the CHN6-CUG road dataset to verify the model’s generalization

performance under complex conditions where urban zones and natural landscapes are intermingled. Table 3 lists the corresponding quantitative measurement results.

Table 3. Comparative Results on the CHN6-CUG Road Dataset

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	IoU (%)
D-LinkNet[6]	96.01	78.53	77.47	77.64	64.02
DSNet[10]	95.48	76.43	74.70	74.96	60.52
BASNet[18]	95.64	75.97	74.97	74.87	60.72
FFNet[19]	95.40	78.39	67.87	72.16	57.24
UNetformer[21]	95.91	78.73	76.78	76.25	63.50
MARENet[12]	95.44	78.14	78.15	78.12	64.71
TSNet	96.10	80.48	79.56	79.37	66.20

From Table 3, it can be seen that TSNet achieves an accuracy of 96.10%, which is 0.66% higher than the suboptimal MARENet, indicating that the model has higher discriminative confidence when dealing with complex background interference. On the Precision metric, TSNet reaches 80.48%, representing a 2.34% improvement over MARENet’s 78.14%, indicating that TSNet can effectively suppress misjudgments in non-road areas, reducing the false positive rate in road extraction. TSNet achieves an F1-score of 79.37%, which is 1.25% higher than the suboptimal

MARENet. TSNet’s IoU reaches 66.20%, a 1.49% improvement over the suboptimal MARENet. Through its unique two-stage processing mechanism, uncertainty-aware learning, and local refinement module, TSNet can effectively handle complex road shapes, shadow occlusions, and spectral confusion in the dataset, maximizing the completeness of the road network while maintaining high precision.

To visually verify the segmentation advantages of TSNet in complex remote sensing scenes, Figure 7 shows qualitative comparisons on the CHN6-CUG road dataset.

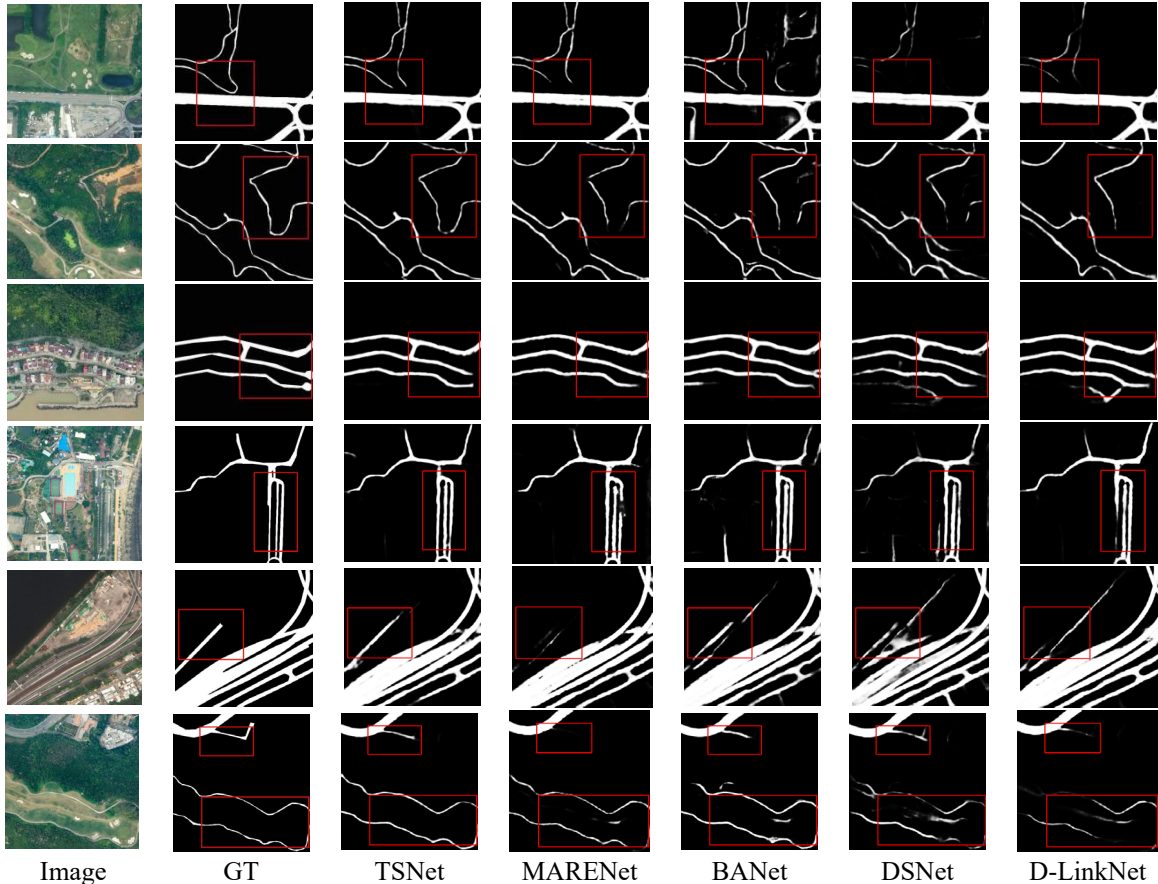


Figure 7. Qualitative visualization of various road extraction algorithms on the CHN6-CUG dataset

As revealed by qualitative comparisons in Figure 7, TSNet is capable of preserving continuous road structures more thoroughly in contrast with alternative algorithms. Obvious fractured road segments can be observed in the outputs

generated by MARENet, BASNet and other comparative models. Benefiting from TSNet’s two-stage optimization scheme, particularly the targeted treatment toward high-uncertainty regions realized by the local refinement

module, the network gains stronger capacity to retain structural information for areas with intricate geometric layouts. It achieves superior recovery of road geometric shapes as well as topological connectivity, which further demonstrates the model's robustness across diverse real-world scenarios.

4. Conclusion

Targeting major technical bottlenecks existing in road-extraction tasks for high-resolution remote-sensing images, including indistinct boundaries, vulnerable fine-grained features and unsatisfactory road continuity, this study develops an uncertainty-guided two-stage segmentation network named TSNet. Taking the CRM module as its feature-extraction backbone, this network innovatively combines uncertainty-aware mechanisms with two-stage progressive optimization strategies. Relying on the joint effects of cross-layer feature enhancement, coarse-segmentation decoding, uncertainty quantification and local refinement operations, this work establishes a processing pipeline following the logic of "global localization-local error correction". Quantitative measurements and qualitative visualization results obtained on public DeepGlobe and CHN6-CUG datasets sufficiently validate that TSNet reaches competitive performance across five key evaluation metrics: Accuracy, Precision, Recall, F1-score and IoU. The proposed model presents favorable robustness for complex urban environments, rural road regions and shadow-occluded zones. Visual comparisons indicate that TSNet can better preserve the topological connectivity of road networks and reconstruct road edge details with higher precision when compared against other mainstream segmentation models. Even so, the model still faces inherent constraints when handling road objects with highly complex topological layouts. Besides, its inference efficiency for large-size images needs further improvement. Subsequent research will concentrate on investigating lightweight two-stage architectures, so as to substantially reduce model parameters and computational overhead without sacrificing segmentation accuracy. Another research direction lies in integrating uncertainty-aware modules into change-detection frameworks, which is expected to realize high-reliability monitoring for dynamic updates of road networks.

References

- [1] Liu, R., Wu, J., Lu, W., Miao, Q., Zhang, H., Liu, X., Lu, Z., & Li, L. (2024). A review of deep learning-based methods for road extraction from high-resolution remote sensing images. *Remote Sensing*, 16(12), 2056. <https://doi.org/10.3390/rs16122056>.
- [2] Abdollahi, A., Pradhan, B., Shukla, N., Chakraborty, S., & Alamri, A. (2022). Deep learning approaches applied to remote sensing datasets for road extraction: A state-of-the-art review. *IEEE Access*, 10, 45205–45239. <https://doi.org/10.1109/ACCESS.2022.3169872>
- [3] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (pp. 234–241). Springer International Publishing. https://doi.org/10.1007/978-3-319-24574-4_28
- [4] Yang, X., Zhang, Y. M., Wang, L., & Zhao, J. H. (2022). Road extraction from high-resolution remote sensing images based on improved U-Net. *Journal of Geomatics*, 47(3), 1–6.
- [5] Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 801–818). https://doi.org/10.1007/978-3-030-01234-2_49
- [6] Zhou, L., Zhang, C., & Wu, M. (2018). D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 182–186).
- [7] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 10012–10022). <https://doi.org/10.1109/ICCV48922.2021.00986>
- [8] Xie, E., Wang, W., Yu, Z., Anandkumar, M., Alvarez, J. M., & Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34, 12077–12090.
- [9] Tao, J., Chen, Z., Sun, Z. C., Guo, H. D., Zeng, B. L., Yu, Z. B., Wang, Y. L., He, Z. Q., & Lei, X. Q. (2023). SEG-Road: A segmentation network for road extraction based on transformer and CNN with connectivity structures. *Remote Sensing*, 15(6), 1602. <https://doi.org/10.3390/rs15061602>
- [10] Guo, Z., Bian, L., Wei, H., Li, J., Ni, H., & Huang, X. (2024). DSNet: A novel way to use atrous convolutions in semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 35, 3679–3692. <https://doi.org/10.1109/TCSVT.2024.3401821>
- [11] Yang, Z., Bi, F., Hou, X., Zhou, D., & Wang, Y. (2024). DDRNet: Dual-domain refinement network for remote sensing image semantic segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17, 20177–20189. <https://doi.org/10.1109/JSTARS.2024.3422347>
- [12] Liu, J., & Huang, C. (2024). Multi-scale adaptive road extraction network for high-resolution remote sensing images. *Frontiers in Computational Intelligence and Systems*, 10(3), 101–106.
- [13] He, F., Liu, S., Liu, S., Jin, Y., Xie, H., & Tong, X. (2025). HDRoad: An encoder-decoder architecture with hybrid attention and directional prior for efficient road extraction from remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 227, 251–264. <https://doi.org/10.1016/j.isprsjprs.2025.06.014>
- [14] Ruby, U., & Yendapalli, V. (2020). Binary cross entropy with deep learning technique for image classification. *International Journal of Advanced Trends in Computer Science and Engineering*, 9(10).
- [15] Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., & Cardoso, M. J. (2017). Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (pp. 240–248). Springer International Publishing. https://doi.org/10.1007/978-3-319-67558-9_28
- [16] Yu, Y., Wang, C., Fu, Q., Kou, R., Wu, W., & Liu, T. (2023). Survey of evaluation metrics and methods for semantic segmentation. *Computer Engineering and Applications*, 59(6), 57–69.
- [17] Li, R., Wang, L., Zhang, C., Duan, C., & Zheng, S. (2022). A2-FPN for semantic segmentation of fine-resolution remotely

- sensed images. *International Journal of Remote Sensing*, 43(3), 1131–1155. <https://doi.org/10.1080/01431161.2021.2022775>
- [18] Bo, W., Liu, J., Fan, X., Tjahjadi, T., Ye, Q., & Fu, L. (2022). BASNet: Burned area segmentation network for real-time detection of damage maps in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–13. <https://doi.org/10.1109/TGRS.2022.3141203>
- [19] Mehta, D., Sklar, A., Ben Yahia, H., Borse, S., Porikli, F., Habibian, A., & Blankevoort, T. (2022). Simple and efficient architectures for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2628–2636). IEEE.
- [20] Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., & Iwamoto, Y. (2020). Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1055–1059). IEEE. <https://doi.org/10.1109/ICASSP40776.2020.9053405>
- [21] Wang, L., Li, R., Zhang, C., Fang, S., Duan, C., Meng, X., & Atkinson, P. M. (2022). UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 190–214). IEEE. <https://doi.org/10.1109/ICASSP43922.2022.9746115>
- [22] Yuan, F., Wang, Z. Y., & Lu, K. (2025). DeepLabv3+ remote sensing image road extraction based on multilevel feature fusion. *Laser Journal*, 46(7), 93–100.