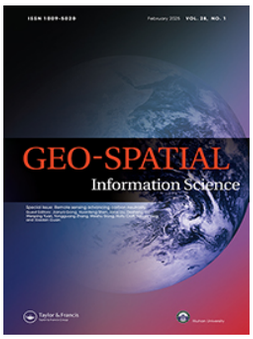

Geo-spatial Information Science

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tgsi20

Narrow road extraction from high-resolution remote sensing images: SWGE-Net and MSIF-Net

Zhebin Zhao, Wu Chen, San Jiang, Yaxin Li & Jingxian Wang

To cite this article: Zhebin Zhao, Wu Chen, San Jiang, Yaxin Li & Jingxian Wang (20 Sep 2024): Narrow road extraction from high-resolution remote sensing images: SWGE-Net and MSIF-Net, Geo-spatial Information Science, DOI: 10.1080/10095020.2024.2405017

To link to this article: https://doi.org/10.1080/10095020.2024.2405017

GEO-SPATIAL INFORMATION SCIENCE
https://doi.org/10.1080/10095020.2024.2405017

Narrow road extraction from high-resolution remote sensing images: SWGE-Net and MSIF-Net

Zhebin Zhao[a], Wu Chen[a], San Jiang[b], Yaxin Li[a,c] and Jingxian Wang[a]

[a]Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong, China; [b]School of Computer Science, China University of Geosciences, Wuhan, China; [c]Algorithm Development Department, Micro Dimension Technology Limited, Hong Kong, China

ABSTRACT

Accurate and complete road network extraction plays a critical role in urban planning, street navigation, and emergency response. At present, narrow roads are a main feature in most public road datasets. However, the continuity and boundary completeness of the extraction results for these narrow roads are relatively poor, due to their varied shapes, uneven spatial distribution, and the presence of various interfering elements. To address these issues, this study introduces a novel network, the Self-weighted Global Context Road Extraction Network (SWGE-Net), which integrates a dilate block and an improved coordinate attention mechanism to effectively capture the complex details and spatial information of narrow roads. Furthermore, most public road training datasets often lack labels for very narrow roads, this omission leads to poor extraction results for these roads in test datasets. In order to further improve the extraction capability for unlabeled, extremely narrow roads, this study introduces another network called the Multi-scale Information Fusion Road Extraction Network (MSIF-Net), which uses the same encoders as SWGE-Net and has a special module for merging information at different scales. This module, with a dilate block and pyramid pooling-based decoder, makes the network better at recognizing and combining features of different sizes. Experimental results indicate that SWGE-Net outperforms the baseline network with road IoU scores of 71.57% and 60.67% on the DeepGlobe and CHN6-CUG road datasets, respectively an improvement of 18.51% and 5.40%. Meanwhile, MSIF-Net not only exceeds the baseline in road IoU scores for both datasets, but also achieves the best performance in extracting unlabeled, extremely narrow roads in qualitative experiments.

ARTICLE HISTORY
Received 12 December 2023
Accepted 11 September 2024

KEYWORDS
Road extraction; high resolution remote sensing images; deep learning

1. Introduction

Road extraction identifies road networks from remote sensing images. This process is essential for urban planning, navigation, military uses, and emergency rescue (Wang et al. 2016). While a well-defined road has clear boundaries and a uniform structure, the complexity of road structures and the details in remote sensing images pose significant challenges (Dewangan and Sahu 2023; Miao et al. 2015). With advances in onboard hardware and higher resolution images, we can now capture detailed features of road targets. However, complex environments and road diversity still present major challenges. Figure 1 illustrates key challenges in extracting narrow roads, affecting their continuity and integrity: (a) Resolution discrepancies can lead to lost road details, with low-resolution images possibly omitting narrow roads. (b) Obstructions from trees, buildings, and other objects can obscure roads, complicating their detection. (c) Atmospheric interference may blur road edges and reduce contrast, hindering road extraction. (d) Complex backgrounds such as mountains, forests, or urban structures challenge the distinction of roads. (e) Lighting conditions and shadows can drastically affect road visibility, with intense sunlight or weak lighting making it difficult to discern road details.All these factors can negatively impact the continuity and integrity of narrow road extraction.

To address the above issues, many researchers have conducted in-depth studies, proposing techniques like the SIIS, semantic segmentation, edge detection, and the CRAE-Net to enhance road extraction accuracy and continuity (Li, Gao, and Xu 2020; Li et al. 2022; Tao et al. 2019). Despite these advancements, the persistent high omission rates for narrow roads underscore the ongoing need for advancements.

In response, we have developed the Self-Weighted Global Context Road Extraction Network (SWGE-Net), which integrates dilation blocks with enhanced coordinate attention modules tailored to meet the unique demands of road extraction tasks. This network utilizes dilated blocks with varying convolution layers to expand its receptive field without complicating the model, significantly enhancing the accuracy and continuity of

CONTACT Yaxin Li yaxin-roan.li@polyu.edu.hk

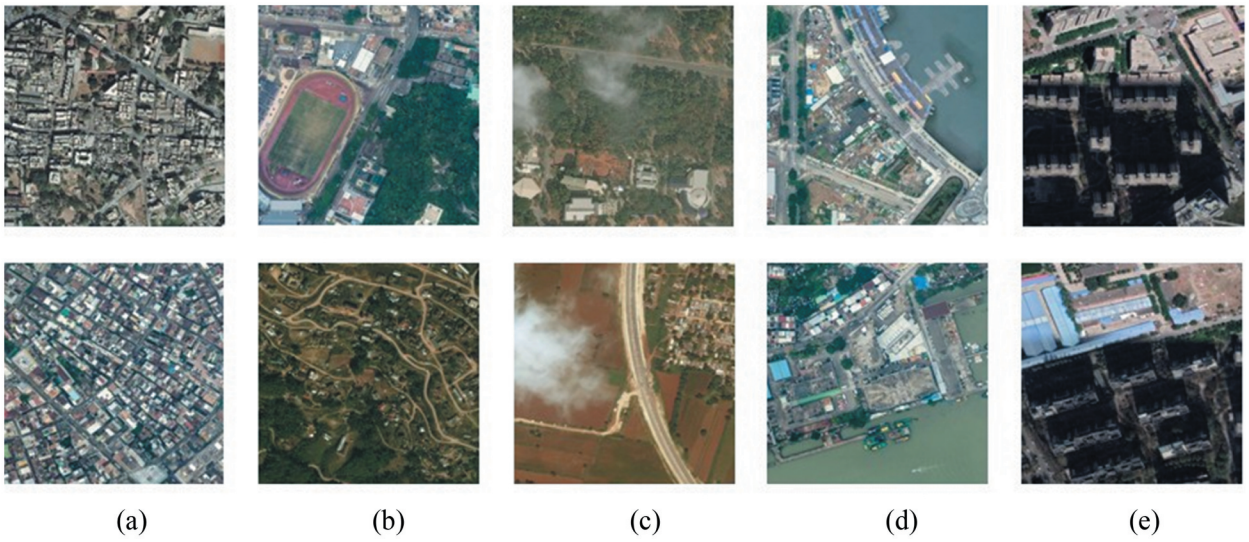

Figure 1. Highlights the principal challenges in road extraction tasks: (a) resolution discrepancies that result in detail loss; (b) obstructions by external objects such as trees and buildings; (c) atmospheric interference causing image noise; (d) complex backgrounds that obscure roads; (e) effects of lighting and shadows on road visibility.

narrow road extraction. Additionally, the integration of a coordinate attention module allows for precise localization of features within specific areas, significantly improving the network's ability to accurately delineate road boundaries. Furthermore, to enhance the detection and extraction of unlabeled, extremely narrow roads, we introduced the Multi-scale Information Fusion Road Extraction Network (MSIF-Net). This network smartly combines dilated blocks with a decoder based on pyramid pooling to manage features at various scales, significantly boosting the model's ability to detect and extract extremely narrow roads. In summary, the main contributions of this study are as follows:

(1) We proposed a Self-Weighted Global Context Road Extraction Network that coupled dilated convolution modules with coordinate attention modules, paying simultaneous attention to local details and global context information. This enhanced the model's ability to infer correlations between homogeneous road objects and improved the continuity and boundary completeness of the narrow road extraction results.
(2) We proposed a Multi-scale Information Fusion road extraction network. This network incorporates a decoder designed with pyramid pooling and jump connections to capture multi-scale contextual information. This enhances the depiction of unlabeled, extremely narrow roads and their relationship with their surroundings. By integrating the expansion block with this decoder, we created a multi-scale information fusion module that improved the extraction of unlabeled, extremely narrow roads.
(3) We employed the DeepGlobe and CHN6-CUG road datasets to evaluate our model's ability for extracting narrow roads and their adaptability across various contexts. Performance was assessed using a confusion matrix and four metrics: precision, recall, F1-score, and IoU. Both quantitative and qualitative analysis indicate that our model significantly outperforms other methods in terms of performance. Furthermore, ablation studies have also validated the effectiveness of the enhancement module we proposed.

The rest of this paper is structured as follows. Section 2 introduces the related works. Section 3 describes the proposed algorithm. Section 4 presents the datasets used, as well as the results and analysis of our experiments. Section 5 draws the conclusions of this study.

2. Related works

Road extraction from remote sensing images initially relied on manual annotation, which was time-consuming and subjective. The introduction of computer vision automated some processes but still had limitations. The shift to deep learning technologies, particularly Convolutional Neural Networks, improved feature learning and handling of complex scenarios. In recent years, the focus of road extraction tasks based on high-resolution remote sensing images has shifted from applying various models to optimizing the network structure for the insurmountable problems of road extraction (Zhang et al. 2023).

2.1. Traditional road extraction methods

Traditional road extraction methodologies, which predominantly utilize spectrum, texture, and geometric

shapes, are bifurcated into two primary analytical scales: pixel-based and object-based methods. Each of these methodologies boasts unique advantages, yet also exhibits inherent limitations.

Pixel-based methods, for instance, emphasize the extraction of spectral and texture features at the individual pixel level. The seminal work of Abdollahi, Bakhtiari, and Nejad (2018) effectively utilized a fusion of Support Vector Machine (SVM) and Level Set (LS) algorithms to extract roads from Google Earth images. Concurrently, the research by Li, Hu, and Ai (2018) proposed an innovative, unsupervised road detection technique that leverages a Gaussian mixture model along with object-based features across a series of processing stages. However, despite their efficacy, these pixel-based techniques may occasionally lead to inaccuracies by the faults of classify objects that bear resemblances to roads.

In stark contrast, object-based techniques treat road entities as comprehensive, interconnected entities rather than isolated pixels. This holistic approach provides robust resistance against common issues such as spectral outliers and salt and pepper noise. For example, Ding et al. (2016) proposed a road extraction methodology based on direction consistency, whereas Maboudi et al. (2018) introduced innovatively a technique that melds ant colony optimization, fuzzy logic systems, and object-based image analysis. Further, Huang and Zhang (2009) pursued a novel extraction technique that leverages the multi-scale structural characteristics inherent in objects.

Despite the relative success of both pixel-based and object-based techniques in extracting roads with clear boundaries in non-complex backgrounds, these methodologies grapple with more intricate situations. Both techniques are susceptible to various forms of noise and exhibit significant performance degradation under complex backgrounds. This underscores the necessity for the development and implementation of more sophisticated, robust techniques for road extraction

2.2. Deep learning based road extraction methods

In the previous section, we saw that traditional road extraction methods work well with roads that have clear boundaries but struggle with more complex environments and narrower roads. They also require a lot of manual work and don't handle large datasets effectively. This sets the stage for a more sophisticated approach. In the upcoming discussion, we'll look at how deep learning-based road extraction methods are gaining ground. These techniques are excellent at interpreting complex data patterns, which not only improves the accuracy of road detection but also cuts down on the need for manual tuning and feature selection. They are particularly good at extracting complex roads, making them a great tool for automating the extraction of roads.

Deep learning technologies offer potential solutions for road extraction from remote sensing images (Hou, Zhou, and Feng 2021; Shamsolmoali et al. 2020; Wei and Ji 2021). This is due to their ability to learn complex patterns in data, adapt to varying conditions, and handle large datasets. They can predict outcomes, like road locations, directly from raw images, bypassing manual parameter adjustments across multiple steps (Abdollahi et al. 2020; Chen et al. 2018; Lian et al. 2020).

Fully Convolutional Networks (FCNs) have been trans-formative in pixel-level image classification, especially in semantic segmentation tasks like road extraction (Long, Shelhamer, and Darrell 2015). However, FCNs face limitations, such as noise introduction and reduced spatial context during upsampling (Badrinarayanan, Kendall, and Cipolla 2017). To address these, FCN variants like DeepLab (Chen et al. 2014), U-Net (Ronneberger, Fischer, and Brox 2015), and SegNet have adopted an encoder-decoder architecture, which effectively maps low-resolution feature maps back to the full input image size (Badrinarayanan, Kendall, and Cipolla 2017). Studies have illustrated the utility of this architecture in road extraction from remote sensing images, with various models demonstrating superior performance (Máttyus, Luo, and Urtasun 2017; Zhang, Liu, and Wang 2018). Xu et al. (2018) proposed a network for extracting roads from aerial images. Their network design combines local and global information, focusing on the local texture and overall morphological structure of roads, respectively, to enhance the perception and aggregation of contextual information. Mosinska et al. (2018) proposed an iterative improvement method for topology extraction based on U-Net. They proposed a new topological loss term that reduces the topological impact of the prediction error for the road extraction task Gao et al. (2018) proposed the multiple feature pyramid network. They exploited multilevel semantic features of HRSI and designed a unique loss function to address the problem of unbalanced categories. At the same time, the encoder-decoder architecture's success is validated by its performance in road extraction competitions, such as Deepglobe 2018 and SpaceNet Challenge 2018 (Buslaev et al. 2018; Lian et al. 2020; Zhou, Zhang, and Wu 2018). These architectures have proven effective in road extraction studies, achieving end to end pixel-level road extraction by reducing and then restoring the resolution of the feature image (Abdollahi et al. 2020; Lian et al. 2020). With the rapid development of Transformer, many

Transformer-based road extraction algorithms have emerged. Zhang, Sun, and Liu (2022) proposed a dual-resolution road segmentation network with a features fusion module for road extraction tasks. Experiments using the Massachusetts dataset and DeepGlobe dataset showed that their proposed network performed excellently (Zhang, Sun, and Liu 2022). Tao et al. (2023) designed a road extraction and segmentation network based on Transformer and CNN with connection structure, which successfully improved the connectivity of road extraction results. Chen et al. (2023) also proposed a dual path extraction network based on CNN and Transformer, which combines local and global features to fully extract the semantic information of the road, effectively improving the integrity of the road extraction results.

In the road extraction task based on high-resolution remote sensing images, the road area usually only accounts for a small part of the entire remote sensing image, and even the width of the road is only a few consecutive pixels, which is different from most semantic segmentation tasks, which gives semantic segmentation The Internet presents huge challenges. At the same time, roads are usually continuous, which means they have strong context dependence, which makes complete extraction of roads more difficult. Most of CNN methods have struggled with capturing long range dependencies in data due to their local receptive fields and in-variance of translation (Dewangan, Sahu, and Arya 2024). Therefore, we need to optimize the road extraction network structure in a targeted manner to overcome these problems (Hinz and Baumgartner 2000). Current studies often utilize attention mechanisms and multi-scale features for this purpose (Jie et al. 2022). Attention mechanisms refine road feature representation and are categorized into channel attention and spatial attention (Guo et al. 2022). Channel attention involves assigning weights to channel dimensions in the feature map, emphasizing crucial road feature channels (Zhang, Sun, and Liu 2022). Spatial attention focuses on spatial locations within the road feature map, prioritizing beneficial regions through learned location weights, and improving road perception and overall image comprehension (Zhang, Sun, and Liu 2022). Li et al. 2020 proposed a semantic segmentation network called SCAttNet, which adaptively refines features by integrating spatial and channel attention modules, and it makes full use of the rich spatial and semantic information in remotely sensed images to improve road extraction. Zhao et al. (2023) added a striped positional attention mechanism between each residual block of their network and also constructed a pyramid expansion module with striped convolution and attention mechanism between the encoder and decoder in order to enhance the coherent semantic information of roads to improve the results of road extraction. Dai, Zhang, and Zhang (2023) proposed road augmented deformable attention network (RADANet)) to learn the long range dependence of specific road pixels. They designed a road augmentation module (RAM) and a deformable attention module (DAM), where the deformable attention module (DAM) combines the sparse sampling capability of deformable convolution with a spatial self-attention mechanism to extract more specific road features. However, there are limitations, such as spatial attention's potential inadequacy in modeling necessary remote dependencies, especially for high-resolution feature maps. Multi-scale feature techniques, including image pyramids, pyramid pooling, skip connections, and dilated convolutions, efficiently capture local context and are extensively applied in road extraction tasks. For example, Zhu et al. (2021) have proposed customized multi-scale modules to enhance accurate road edge perception and feature representation. Du et al. (2023) developed the MSI-guided Segmentation Network for road extraction, merging multi-spectral and neural network strengths to enhance remote sensing image analysis. Their network integrates multi-scale information using LSFF and GASF modules and employs structural loss and channel attention, outperforming existing methods. Although these multi-scale modules are effective in capturing local context, there is still room for improvement in terms of their interaction with feature learning in the encoder. This interaction plays a crucial role in enhancing the model's overall feature representation ability for road extraction. This means that the problems of low continuity and boundary completeness for narrow road extraction, as well as the difficulty of unlabeled, extremely narrow roads to be extracted, remain unavoidable.

In conclusion, while deep learning has significantly advanced the field of road extraction, several pervasive challenges remain. These include difficulties in distinguishing roads from complex backgrounds, handling the variability in road appearances, addressing occlusions and inconsistent lighting conditions, and managing scale differences. Moreover, the prevalent issues of data imbalance, low-resolution imagery, and the scarcity of high-quality labels continue to impede model performance. Generalization across diverse terrains and real-time processing also present substantial hurdles. Finally, the integration of multi-source data and the timely detection of changes in road networks are areas that necessitate further research. Overcoming these obstacles is crucial for the development of robust, accurate, and efficient road extraction systems, and represents a dynamic and critical area of ongoing research in the remote sensing community.

3. Research methodology

In the present study, we propose two innovative encoder-decoder architecture networks, designed specifically for the extraction of roads from remote sensing imagery: the SWGE-Net and the MSIF-Net. Notably, SWGE-Net is purposefully constructed to confront the challenges of inadequate continuity and boundary integrity that are frequently encountered by extant methods during the delineation of narrow road boundaries. Concurrently, MSIF-Net has been designed with a focus on enhancing the capability to detect and extract roads of extremely poor quality.

3.1. Overall structure of SWGE-Net and MSIF-Net

SWGE-Net and MSIF-Net are both designed based on the D-LinkNet network. D-LinkNet is a deep learning model for road extraction tasks. It improves on the LinkNet structure, adds Dilated Convolutions Modules, and expands the receptive field of the network by introducing dilated convolutions. This enables the model to capture broader contextual information without adding additional computational burden. D-LinkNet is designed to be particularly suitable for addressing the challenges in road extraction and remote sensing image analysis, as it can effectively handle scale changes in images and performs well in extracting small and complex road structures. The details of D-LinkNet are shown in Figure 2. It has generally achieved excellent results across various road datasets. However, it also has the following problems:

(1) Lack of spatial attention mechanism: D-LinkNet may not have fully utilized the spatial attention mechanism. This mechanism can help the model focus on key areas in the image, such as narrow roads. If this mechanism is not fully used, the model may miss these important details.This leads to poor continuity and low boundary integrity of road extraction results.
(2) Limitations of multi-scale feature representation: D-LinkNet uses dilated convolutions to capture road features, which is very important for capturing road features of different scales and shapes. However, for extremely narrow roads, this multi-scale representation may not be able to capture these tiny features with sufficient precision.This has resulted in some extremely narrow roads that are extremely difficult to extract.

In response to the lack of spatial attention mechanisms, we designed SWGE-Net. SWGE-Net comprises three components: an encoder, a Self-weighted global feature extraction module(SWGE), and a decoder. The SWGE module mainly consists of two parts: the Dilated Convolution module and an enhanced Coordinate Attention module (Hou, Zhou, and Feng 2021). Dilated Convolution expands the receptive field by introducing dilation in the convolutional kernel, enabling the capture of broader contextual information. Coordinate Attention is an attention mechanism utilized for image processing tasks, enhancing the importance of different positions in the image by learning attention weights on spatial coordinates.

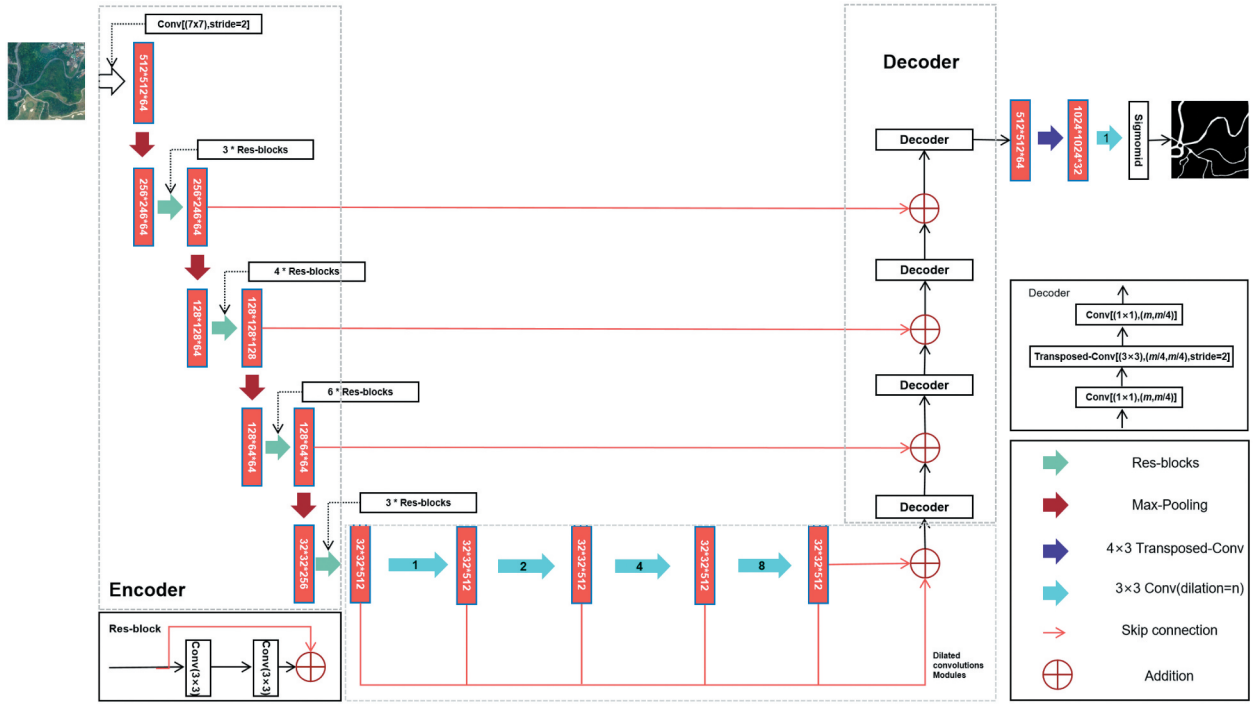

Figure 2. D-LinkNet.

This combination effectively utilizes the expanded receptive field properties of the dilated convolution and intensifies focus on critical locations via the coordinate attention mechanism. Thereby improving the continuity and boundary integrity of the results from narrow road extraction. Figure 3 presents the primary architecture of this method.

In response to the limitations of multi-scale feature representation, we designed MSIF-Net. The multi-scale information fusion module of MSIF-Net combines Dilated Convolution with a decoder based on Pyramid Pooling. Dilated Convolution expands the receptive field by introducing dilation in the convolutional kernel, capturing a broader contextual scope. The decoder based on Pyramid Pooling conducts feature pooling at different scales, capturing and integrating richer multi-scale contextual information (Zhao et al. 2017). This augmentation enhances the model's robustness and generalization ability, thereby elevating the capability to extract extremely narrow roads, as shown in Figure 4.

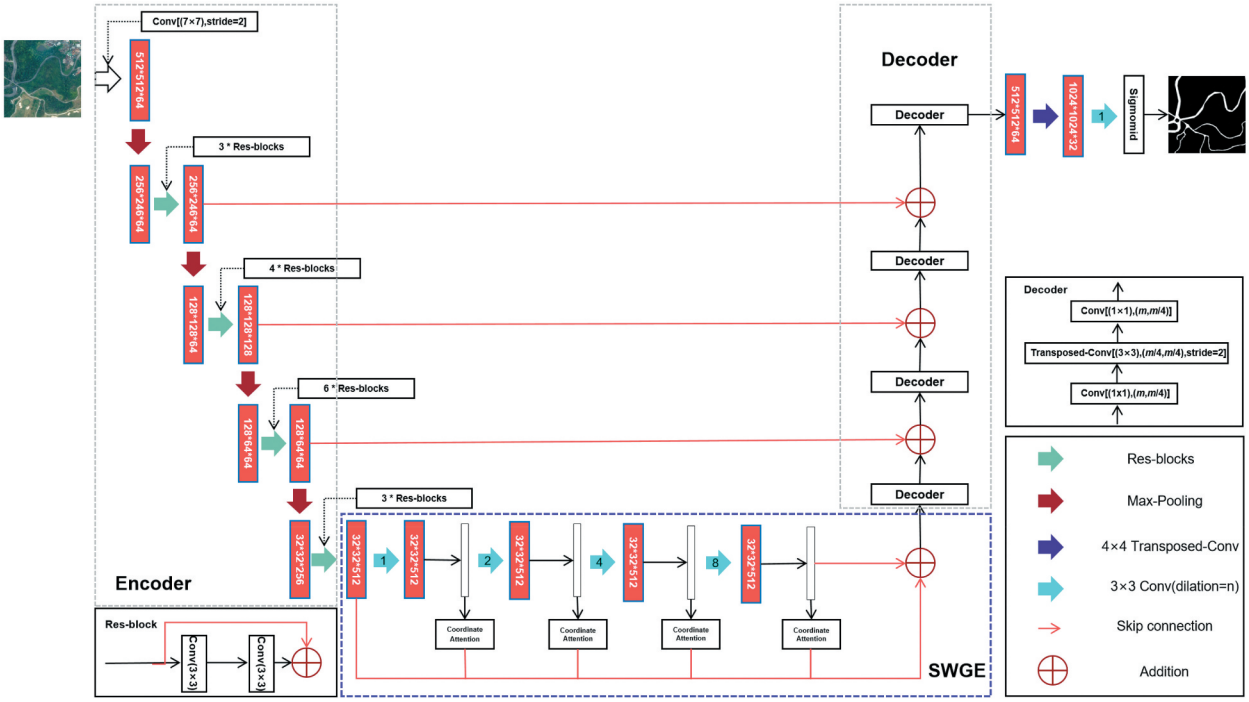

Figure 3. Self-weighted global context road extraction network (SWGE-Net).

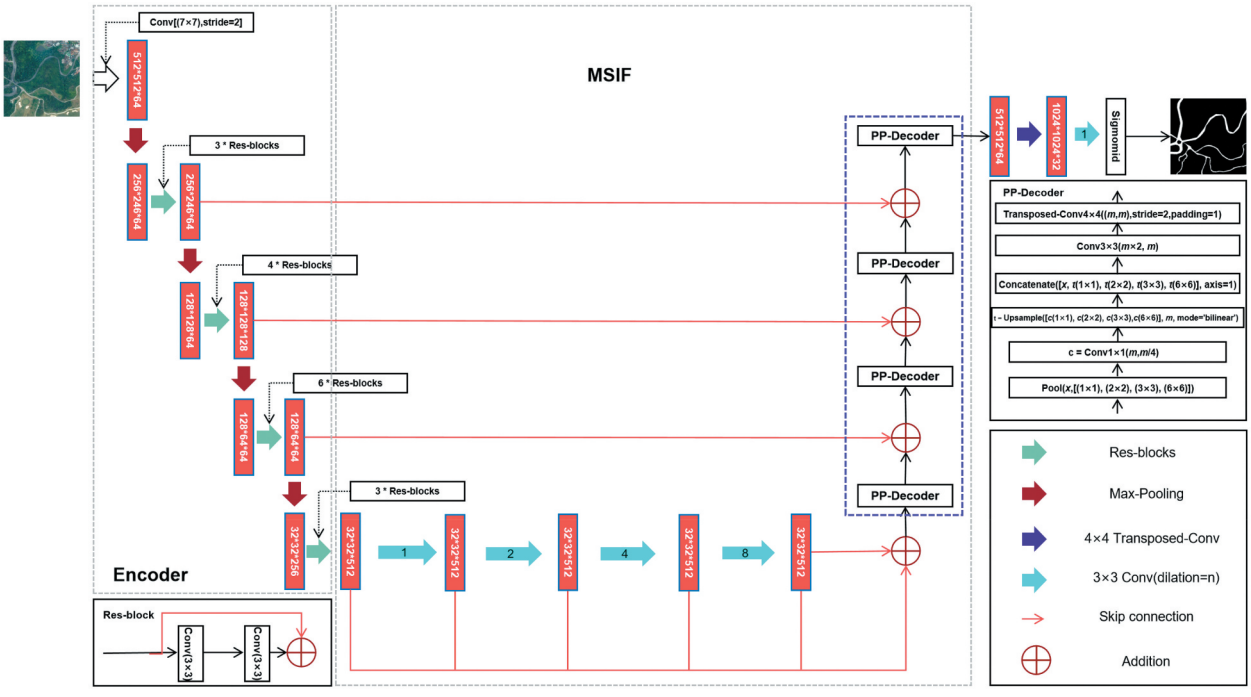

Figure 4. Multi-scale information fusion road extraction network (MSIF-Net).

3.2. Self-weighted global feature extraction module

The Self-weighted Global Feature Extraction (SWGE) module is designed to improve the continuity and boundary integrity of the extraction results for narrow roads. It consists of a combination of a dilated convolution module and a coordinate attention module, and the structure is shown in Figure 5. The combined use of these two modules can achieve complementary advantages in feature extraction. In dilated convolution, by inserting dilation of different sizes between elements of standard convolution kernels, the receptive field of the network can be expanded without adding additional computational costs or reducing image resolution (Zhou, Zhang, and Wu 2018). This allows the model to perform without losing detail information (Zhou, Zhang, and Wu 2018). Under the premise, capture a wider range of contextual information. These feature maps then enter the coordinate attention module for further processing. This module generates attention maps to weight the feature maps by learning the distribution of the feature maps in the horizontal direction (X) and vertical direction (Y), effectively emphasizing important road information and Suppress unnecessary additional information (Hou, Zhou, and Feng 2021). This re-calibration not only occurs at a single scale, but is repeated across features at multiple different scales to ensure that features captured from different levels are optimally tuned (Hou, Zhou, and Feng 2021). Through this tight coupling, the model can process features more granular at each level, improving the continuity and boundary integrity of the roads in its extracted images, especially for narrow roads.

As shown in Figure 5, if the Dilated Convolution layers are stacked with dilation rates of 1, 2, 4, and 8, the receptive fields of each layer will be 3, 7, 15, and 31, respectively. The encoder section consists of 5 downsampling layers. Taking an image of size 1024×1024 as an example, the output feature map will be 32×32. In this scenario, the feature points on the last central layer will perceive a feature map of size 32×32. Consequently, the feature points on the last central layer will observe 31×31 points on the first central feature map, covering the primary portion of the initial central feature map.

As shown in Figure 6, an enhanced coordinate attention mechanism is incorporated after each Dilated Convolution layer simultaneously. The coordinate attention module can efficiently capture global attention in both feature channels and spatial positions. In comparison with other attention mechanisms, it possesses lower computational and parameter overheads. This mechanism, acting as a lightweight module, integrates seamlessly with the Dilated Convolution layers.

In most cases, global average pooling is commonly employed for channel attention. However, it compresses global spatial information into channel descriptors, making it challenging to preserve location details. In tasks such as road extraction, features with

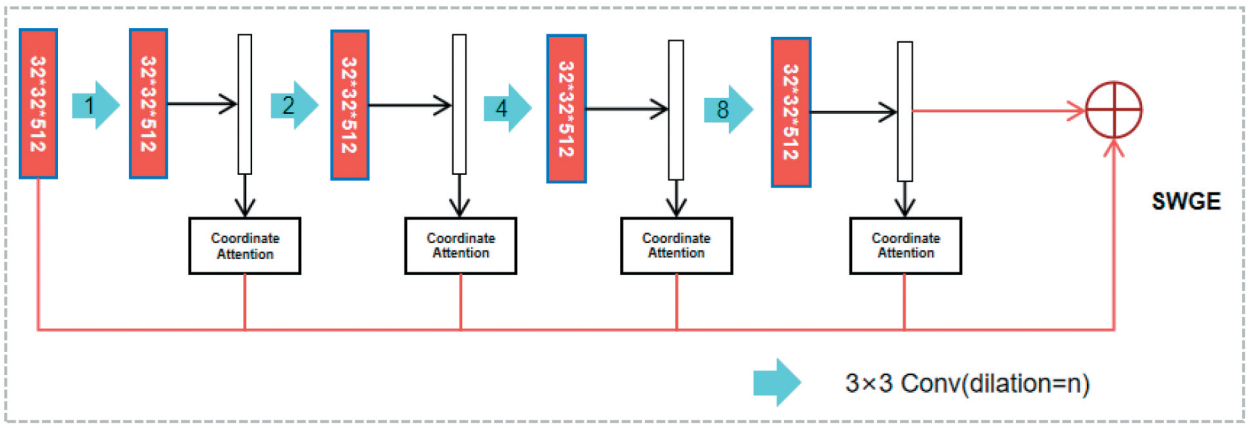

Figure 5. Self-weighted global feature extraction module.

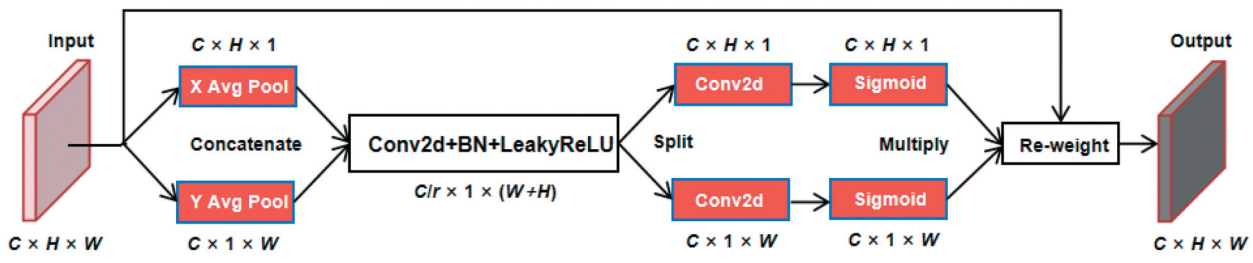

Figure 6. An enhanced coordinate attention mechanism.

preserved spatial positional information are crucial for capturing road spatial structures. Therefore, it is crucial to retain accurate spatial information of road features during the feature compression process while capturing global feature information.

During the embedding of coordinate information, two 1D global average pooling operations substitute the 2D global average pooling operation. This approach better captures long range spatial interactions and precise positional information, enabling the attention module to more accurately capture relationships between features. For the input feature map, for each channel, two 1D average pooling kernels, namely (1, W) and (H, 1), are applied along the horizontal and vertical dimensions, respectively. After information compression, two feature tensors, denoted as and, are obtained, incorporating spatial information from different positions. The output of the c-th channel at height "h" or width "w" can be expressed as follows:

$$f_c^h(h) = \frac{1}{W}\sum_{0\le i<W} x_c(h,i) \tag{1}$$

$$f_c^w(w) = \frac{1}{H}\sum_{0\le j<H} x_c(j,w) \tag{2}$$

The equations integrate features from different directions, thereby generating a pair of directional feature maps. Compared to the global pooling compression method, this approach allows the attention module to capture long range relationships in one direction while simultaneously preserving spatial information in another, thus enabling the network to localize targets more accurately.

During the coordinate attention generation process, the model can leverage contextual information from various directions to locate the road regions of interest and generate effective spatial and channel attention weights, thereby indirectly enhancing occluded road features. Initially, feature tensors from horizontal and vertical directions are concatenated, forming a new feature tensor denoted as $f \in R^{C\times 1\times (W+H)}$. Subsequently, the feature tensor undergoes transformation through a shared 1×1 standard convolution, generating a dimensional reduced feature tensor $f \in R^{C/r\times 1\times (W+H)}$, where r represents the down-sampling ratio of the channel dimension. This module then processes the tensor through batch normalization layers and nonlinear activation layers, separating the feature tensor F into two directional feature tensors, $F^X \in R^{C/r\times H\times 1}$ and $F^Y \in R^{C/r\times 1\times W}$. Then, two 1×1 standard convolutions are applied to compute attention tensors $Z^x \in R^{C\times H\times 1}$and $Z^y \in R^{C\times 1\times W}$ in two directions. Ultimately, the attention tensors are normalized using the sigmoid function and constrained within a range of 0 to 1. The global attention weight matrix $Z \in R^{C\times H\times W}$ is obtained by matrix multiplication to yield $Z^x$and Z^y, incorporating adaptive channel and spatial dimensions of attention. Subsequently, this module multiplies the attention weight Z with the initial input X, realizing the process of re-weighting process and obtaining the final output $Y \in R^{C\times H\times W}$, thereby optimizing attention.

The activation function in the original coordinate attention mechanism is a customized activation function designed by the authors based on ReLU6, while for the complexity of remote sensing images, the network may receive many negative inputs. The output of the original activation function is 0 when the input is negative, i.e. they cannot learn anymore because their gradient is 0, which may cause some neurons to "die" during the training process, thus losing some critical road information. However, Leaky ReLU is different in that even when it receives negative inputs, the neurons maintain a small positive gradient and thus continue to learn, which makes it very suitable for use in the coordinate attention module of a remote sensing image-based road extraction task. The following equations illustrate the detailed computation process:

$$F = \mathrm{LeakyReLU}(\mathrm{BN}\,(\mathrm{Conv}([f^x, f^y]))) \tag{3}$$

$$Z^x = \mathrm{Sigmoid}\,(\mathrm{Conv}_x(F^x)) \tag{4}$$

$$Z^y = \mathrm{Sigmoid}\,(\mathrm{Conv}_y(F^y)) \tag{5}$$

$$Z = \mathrm{Multiply}\,(Z^x, Z^y) \tag{6}$$

$$Y = X{*}Z \tag{7}$$

$$\mathrm{LeakyReLU}(x) = \begin{cases} x, & if\ x>0 \\ \alpha x, & if\ x\le 0 \end{cases} \tag{8}$$

where Conv() represents the convolution operation, BN () represents the batch normalization operation, LeakyReLU () represents the nonlinear activation function, Multiply() represents the matrix multiplication operation, * represents the element-wise multiplication. α represents a constant smaller than 1, commonly referred to as the "leakage coefficient." In this context, its value is 0.01.

3.3. Multi-scale information fusion module

Multi-Scale Information Fusion Module (MSIF) in order to enhance the ability to extract extremely narrow roads. It consists of a combination of a dilated convolution module and the decoder based on pyramid pooling, and the structure is shown in Figure 7. The clever synergy between the dilated convolution module and the decoder based on pyramid pooling in series particularly enhances the network's ability to identify extremely narrow and unlabeled roads. The

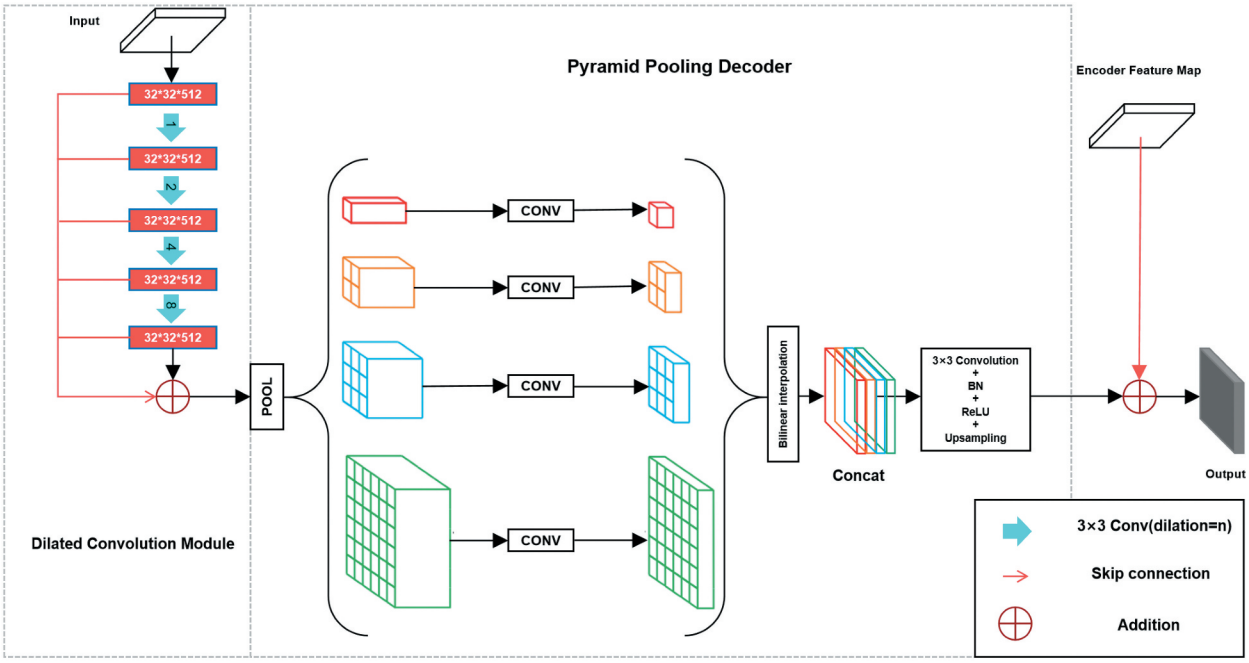

Figure 7. Multi-scale information fusion module.

dilated convolution module effectively expands the size of the model's receptive field by using convolution kernels with different hole rates, thereby capturing richer contextual information without sacrificing resolution (Zhou, Zhang, and Wu 2018). However, this feature is of limited use for discovering and tracking small or obscure road contours, because these road elements often occupy only a few pixels in the image and are difficult to capture. However, inputting the feature map output from the dilated convolution module into a pyramid-pooling decoder can largely compensate for this shortcoming, which aggregates road features at different scales to restore the detailed information and overall structure of the road. This multi-scale feature fusion mechanism provides the model with the ability to understand road shape and width at coarse to fine levels (Shamsolmoali et al. 2020). During the decoding process, the role of the pyramid pooling layer is to unify the features captured from different layers, ensuring as much as possible that even the tiniest roads can be reconstructed and accurately mapped in the final output image. At the same time, the skip connection layer connecting the encoder and the pyramid pooling-based decoder directly transfers the detail information of the shallow layers to the deep layers. This connection not only helps to maintain the details and edge clarity of the roads in the image, but also It promotes the direct reverse flow of gradients, alleviates the problem of gradient disappearance, and ensures effective training of the model (Zhou, Zhang, and Wu 2018). Skip connections are particularly important when extracting extremely small and complex road structures, as they help the model capture and retain subtle feature information that may be lost during the resolution reduction process, identifying and reconstructing roads as accurately as possible (Zhou, Zhang, and Wu 2018).

As shown in Figure 7, the initial part of the MSIF module comprises Dilated Convolution layers, which share the same dilation rates as those established in the Weighted Global Information Extraction module, namely 1, 2, 4, and 8. As a result, the receptive fields for each layer measure 3, 7, 15, and 31, respectively. Taking an image of size 1024×1024 as an example, the output feature map will be 32×32.

The feature map is subsequently fed into the Pyramid Pooling Decoder. Here, within the adaptive max pooling layer, parallel operations are utilized to adjust the output size of the pooling layer, contingent on the specified pool sizes. In the Pyramid Pooling Decoder module, the pool sizes parameter prescribes a list of pool sizes for adaptive max pooling. If pool sizes are (1, 2, 3, 6), then for each pool size, the output size of the adaptive max pooling layer will be adjusted in each spatial dimension. Given an input size of (32, 32, 512), the output size of the adaptive max pooling layer will be adjusted in each dimension based on the specified pool size. The specific output sizes can be calculated as follows: for a pool size of 1×1, the output size is (1, 1, 512); for a pool size of 2×2, the output size is (2, 2, 512); for a pool size of 3×3, the output size is (3, 3, 512); and for a pool size of 6×6, the output size is (6, 6, 512). This occurs as the adaptive max pooling layer auto-adjusts the output size according to both the input size and the specified pool size, while the channel number in each dimension is maintained. Subsequently, through bilinear interpolation, the output size of the adaptive max pooling is adjusted

to the size of the original input feature map. The four feature maps undergo concatenation, followed by a 3×3 convolution, normalization, ReLU activation, and up-sampling to yield the output feature map. Finally, the output feature map is derived by augmenting the feature map that corresponds to the size of the encoder part via skip connections.

The large receptive field of the dilated convolution module and the detailed reconstruction of the pyramid pooling decoder complement each other to form a powerful Multi-Scale Information Fusion Module. This module enables our network to not only identify main road, but also maximize the extraction of extremely narrow and unlabeled roads that are easily overlooked.

4. Experiment and analysis

In this study, we employed two road datasets: DeepGlobe (Demir et al. 2018), covering diverse road types and conditions, and CHN6-CUG (Zhu et al. 2021), focusing on Chinese urban and rural infrastructure. Their combined use allowed a comprehensive assessment of our network's capabilities, particularly for extracting narrow roads and their generalization ability across various environments and styles. We evaluated the model using a confusion matrix and four core metrics: precision, recall, F1 score, and Intersection over Union (IoU). The results showed significant improvements in key performance indicators, notably IoU and F1 scores, compared to other methods. Visualizing the results revealed our model's effectiveness in delivering comprehensive, continuous, and less noisy road extraction, maintaining road continuity and edge integrity. These findings underscore our network's superior performance in complex road extraction tasks, particularly for narrow roads. Figure 8 shows samples from the DeepGlobe dataset, while Figure 9 displays samples from the CHN6-CUG Road dataset.

4.1. Datasets and experimental parameters

To evaluate our road extraction network, we chose the DeepGlobe road dataset, a well-known benchmark that includes a diverse range of road types and conditions found in rural and suburban areas of Thailand, Indonesia, and India. This dataset is especially challenging due to its inclusion of numerous narrow roads, which are critical for depicting complex road infrastructures and ensuring the connectivity of

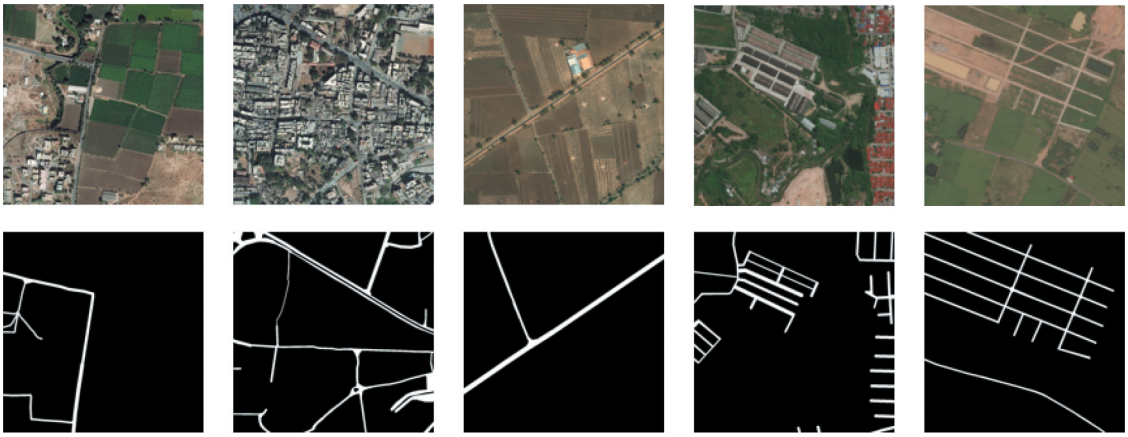

Figure 8. The samples of DeepGlobe Dataset.

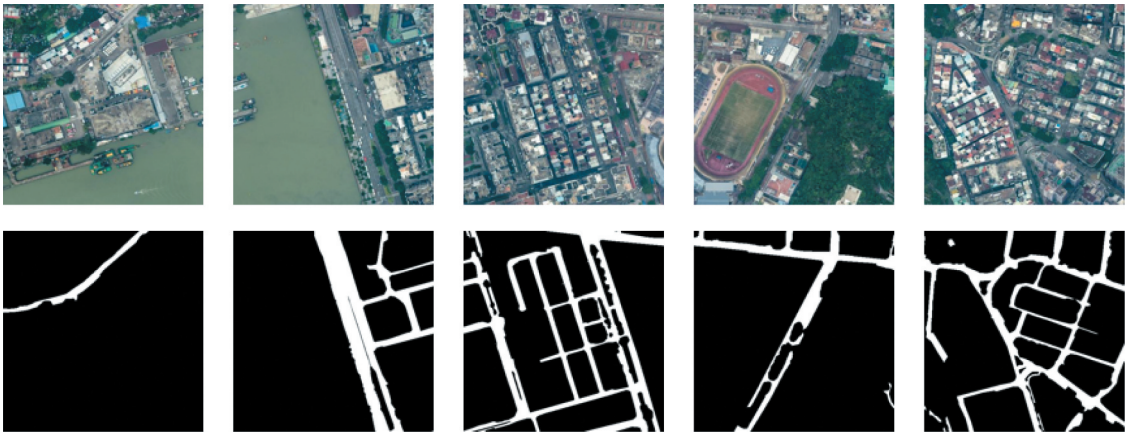

Figure 9. The samples of CHN6-CUG road Dataset.

remote communities. The DeepGlobe dataset encompasses 2220 square kilometers with high-resolution satellite images, and out of its 8570 images, 6226 are accompanied by ground truth data, providing a robust basis for testing our model's scalability and efficiency. We chose a 9:1 training-to-testing ratio to allow extensive learning from the broad array of examples in the training set, which is particularly beneficial for the nuanced detection of narrow roads, while the 10% reserved for testing ensures that the network can generalize well and maintain accuracy when exposed to new data, thereby mitigating the risk of over-fitting and validating the reliability of our road extraction capabilities.

Additionally, we employed the CHN6-CUG dataset, focused on the urban and rural infrastructure of China. This dataset includes intricate urban road layouts and a significant number of narrow roads, typical of rural areas and older urban districts. The presence of these narrow roads introduces additional complexity and reflects the real-world challenges of densely built environments with historical road networks. The CHN6-CUG dataset offers a large-scale, pixel-level collection of very high-resolution satellite images from major Chinese cities, with 4511 annotated images, of which 3608 are used for training and 903 for testing. This dataset's variability, characterized by its diverse road networks within complex urban landscapes, significantly enhances the difficulty of road extraction tasks.

By incorporating both the CHN6-CUG and DeepGlobe datasets in our evaluation, we aim to thoroughly assess our network's capability in extracting narrow roads across varied environments. The CHN6-CUG dataset, with its complex urban road networks and high variability in road width and appearance, tests the network's ability to handle detailed and densely packed road structures. In contrast, the DeepGlobe dataset includes a broader range of road types, from rural to urban, with varying levels of occlusion and environmental conditions, providing a comprehensive challenge for evaluating the network's adaptability and robustness. This approach allows us to gauge the system's performance in a broader set of scenarios, acknowledging that while these datasets enhance the robustness of our validation, they do not encompass all possible road conditions. Our intention is to continuously improve the network's adaptability to diverse environments, with the understanding that additional datasets could further expand the network's generalization and accuracy.

In this study, we conducted comparative experiments using a consistent set of training and testing samples, identical to those used in other relevant models. The experiments employed a consistent set of training and testing samples. The experiments were performed on a computer running the Windows 10 Professional operating system, equipped with a 13th Gen Intel(R) Core(TM) i9-13900KS CPU, 64 GB of RAM, an NVIDIA GeForce RTX 4090 graphics card with 24 GB of VRAM, using CUDA version 11.1 and PyTorch version 1.7.0. We used the ReLU activation function and the Adam optimizer, known for its effective handling of sparse gradients on noisy problems. The loss function combined BCE and Dice loss, optimizing for both accuracy and overlap between the predicted and actual labels. The learning rate was set at 2e-4, with a batch size of 8 to balance between memory usage and model performance. The training process spanned over 200 epochs, allowing sufficient iterations for the networks to converge and learn from the data.

4.2. Evaluation metrics

This paper employs confusion matrices to assess the model performance in binary classification problems. Labels are divided into positive and negative samples, while predictions are categorized as true positive (*TP*) and true negative (*TN*) results, as well as false positive (*FP*) and false negative (*FN*) outcomes. *TP* denotes correct predictions, while *FP* signifies false positives. The performance of the proposed model in road extraction is evaluated using four metrics: precision, recall, F1-score, and Intersection over Union (*IoU*).

4.2.1. Precision

Precision evaluates how many pixels the model correctly predicts as roads. High accuracy means that the model is very accurate at labeling pixels as roads, with very few false positives.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{9}$$

4.2.2. Recall

Recall measures the model's ability to detect all real roads. A high recall rate means that the model can identify most of the real road pixels and rarely misses detection.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{10}$$

4.2.3. F1 score

The F1 score is the harmonic mean of precision and recall, which helps us understand the model's balance between reducing false positives (increasing *precision*) and false negatives (improving *recall*). This is a comprehensive metric that is particularly suitable for evaluating the overall performance of the model in road extraction tasks.

$$\mathrm{F1} = \frac{2TP}{2TP + FP + FN} \tag{11}$$

4.2.4. Intersection over union (IoU)

Intersection over Union is a key indicator for evaluating model performance in segmentation tasks. It calculates the ratio of the overlap area between the predicted road segmentation area and the actual road and their union. A high IoU score indicates that the model spatially accurately identifies and extracts road areas.

$$\mathrm{IoU} = \frac{TP}{FP + TP + FN} \tag{12}$$

By combining these evaluation indicators, we can comprehensively understand the model's performance in the road extraction task, so as to optimize the model to meet the needs of practical applications.

4.3. Experiments on the DeepGlobe dataset

Table 1 lists the qualitative analysis of road extraction on the DeepGlobe dataset. Compared to other methods, our SWGE-Net demonstrates an IoU improvement of 18.51% to 82.57%, and an F1 Score increase of 10.80% to 48.14%. MSIF-Net also shows progress, with an IoU increase of 2.07% to 57.24%, and an F1 Score enhancement of 1.29% to 35.42%. These results highlight the significant advancements of our models over the original ones in key metrics, demonstrating the superior performance of our network in road extraction tasks.

Compared to the common baseline, D-Linknet34, our models demonstrate superior precision in F1 and IoU metrics. Specifically, SWGE-Net, which incorporates a coordinate attention mechanism into D-Linknet34's expansive convolution module, shows an 18.51% increase in IoU and an 10.80% improvement in F1 Score. Moreover, when replacing D-Linknet34's decoder module with a pyramid pooling decoder, SWGE-Net yields a 2.07% increase in IoU and a 1.29% enhancement in F1 Score. These results confirm the advantageous role and effectiveness of the introduced modules.

The integration of coordinate attention and dilated convolution mechanisms within our model has substantially elevated the performance of road extraction, as evidenced by improvements in the IoU and F1 scores. Coordinate attention sharpens the extraction process by highlighting critical road features, enhancing boundary precision, and augmenting the intersection aspect of the IoU. Meanwhile, dilated convolution broadens the model's contextual grasp, capturing extensive spatial information that reduces road fragmentation. This dual enhancement not only fortifies the IoU by refining road delineation but also bolsters the F1 score by increasing TP and minimizing FP, thereby optimizing the results of road segmentation.

Figure 10 shows the qualitative analysis results on the DeepGlobe dataset. The images show the original test set, ground truth, and results of Unet, Unet++, DeeplabV3+, D-LinkNet, SWGE-Net, and MSIF-Net respectively. The test images primarily feature narrow roads.The blue solid lines highlight the most significant parts of the comparison in terms of continuity and boundary completeness of the narrow road extraction results, while the red dashed lines display whether some unlabeled, extremely narrow roads have been successfully extracted.

In the first and fifth rows, the similarity between the road and its background is high, which usually makes narrow road extraction difficult. However, SWGE-Net can effectively capture the continuity and edge details of the road, the performance is largely attributed to the coordinate attention mechanism, which effectively emphasizes important road information and suppresses unnecessary information as much as possible by learning the spatial distribution of feature maps.Thanks to the pyramid pooling-based decoder to aggregate road features at different scales to recover the information and overall structure of the road, MSIF-Net can even extract unlabeled extremely narrow roads. In the second, third, and fourth rows, the presence of numerous trees causes the roads to be obscured, which typically disrupts the continuity and boundary completeness of the road extraction results. However, SWGE-Net maintains a balance of continuity and boundary completeness for narrow roads as much as possible, achieving the best overall effect, this is still a reflection of the effectiveness of the coordinate attention mechanism. MSIF-Net displays the best ability to extract extremely narrow roads.

Table 1. Quantitative evaluation results of different methods on the DeepGlobe dataset.

Model Name	Backbone	IoU(%)	F1-score(%)	Precision(%)	Recall(%)
Unet	None	53.99	70.12	85.32	59.52
Unet++	None	39.20	56.32	84.86	42.15
Deepglobelv3+	Xception	44.23	61.34	90.10	46.49
D-linknet34	ResNet34	60.39	75.30	86.36	66.76
SWGE-Net	ResNet34	**71.57**	**83.43**	84.18	**82.67**
MSIF-Net	ResNet34	61.64	76.27	**87.99**	67.30

In this table, the bold indicates the best result of the objective index in these algorithms.

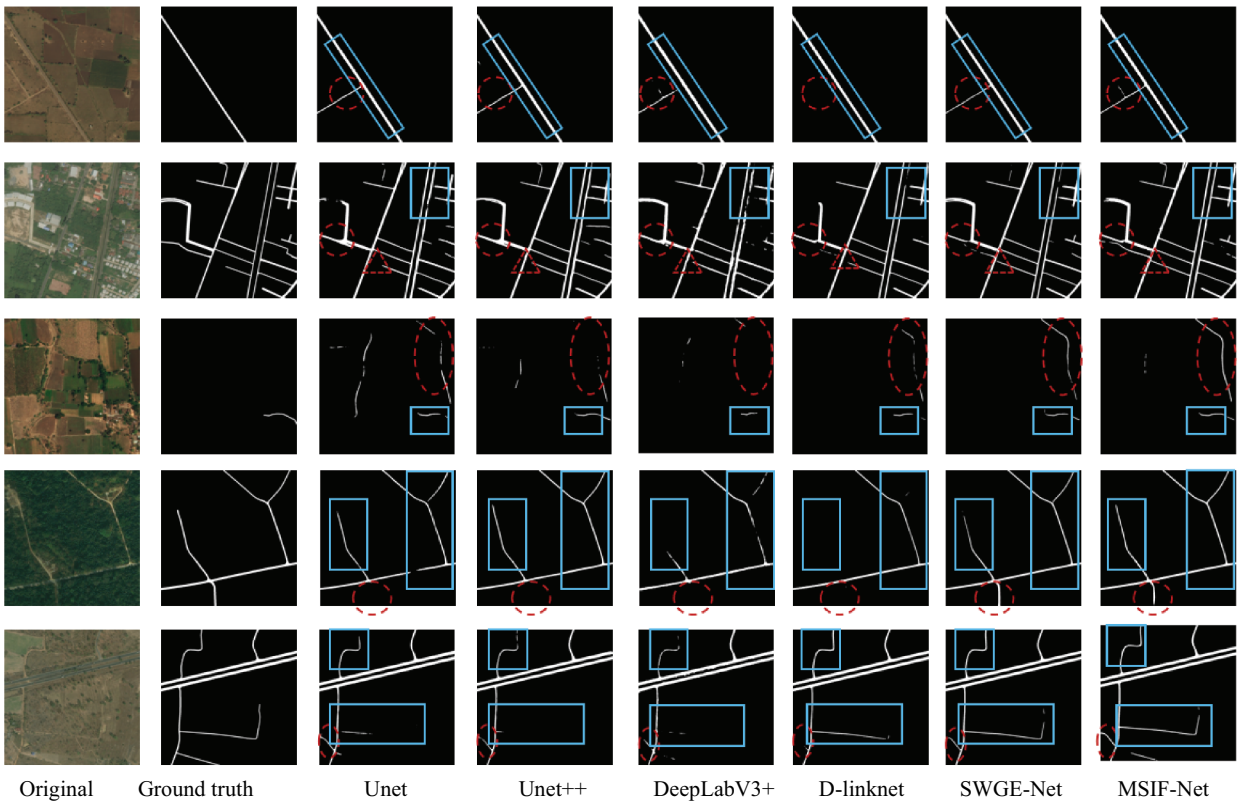

Figure 10. Qualitative evaluation results of different methods on the DeepGlobe dataset.

4.4. Experiments on the CHN6-CUG dataset

We further tested the SWGE-Net and MSIF-Net models on the CHN6-CUG dataset. Table 2 presents the results of the quantitative analysis for road extraction using the CHN6-CUG dataset. It should be noted that the results from the literature (CADUNet, TransUNet, Swin Transformer, SegNet, GCBNet, and CoSwin Transformer) are simply quoted from (Chen, Jiang, and Li 2022; Zhu et al. 2021) and, hence, some metrics and qualitative evaluation results are missing as they were not available.

Among these models, the road IoUs of the proposed models (SWGE-Net and MSIF-Net) are 60.67% and 59.31% respectively. The F1-scores are 74.25% and 72.73% respectively. SWGE-Net and MSIF-Net have the second and fourth performance, respectively, in terms of road IOU, and the second and third respectively in terms of F1-score. Although the performance of our proposed model is not optimal, the results show the effectiveness of our proposed Self-weighted Global Feature Extraction Module and Multi-Scale Information Fusion Module.

Compared to the shared baseline, D-linknet 34, the two proposed models demonstrate higher precision in the comprehensive metrics of F1 and IoU. For instance, the SWGE-Net network exhibits an IoU improvement of 5.40% and an F1 Score increase of 7.20%. Similarly, the MSIF-Net network displays an IoU improvement of 3.04% and an F1 Score increase of 5.01%. These findings affirm the necessity and effectiveness of the introduced modules, suggesting their advantageous role.

The combination of dilated convolutions and pyramid pooling decoder within the MSIF-Net contributes

Table 2. Quantitative evaluation results of different methods on the CHN6-CUG dataset.

Model Name	Backbone	IoU(%)	F1-score(%)	Precision(%)	Recall(%)
Unet	None	48.57	63.77	68.42	59.72
Unet++	None	47.38	63.91	68.33	60.02
Deepglobelv3+	Xception	52.04	65.20	72.24	59.41
D-linknet34	ResNet34	57.56	69.26	72.61	66.21
CADUNet	–	43.47	60.60	68.69	54.21
TransUNet	–	31.74	48.18	69.84	36.78
Swin Transformer	Swin-T	34.10	50.86	78.03	37.72
SegNet	ResNet34	37.24	54.27	62.79	47.78
GCBNet	ResNet34	60.44	72.70	–	–
CoSwin Transformer	ResNet34	**61.28**	**75.99**	**79.75**	72.57
SWGE-Net	ResNet34	60.67	74.25	75.69	**72.86**
MSIF-Net	ResNet34	59.31	72.73	74.71	70.85

In this table, the bold indicates the best result of the objective index in these algorithms.

collaboratively to the significant improvements in both IoU and F1 scores. Dilated convolutions enhance the model's field of view to capture a broad range of contextual details, which, when fused with the multi-scale features from the pyramid pooling, results in a highly detailed and accurate road extraction. This comprehensive feature representation ensures a substantial overlap between the predicted road networks and the ground truth, leading to a higher IoU. Simultaneously, the fusion of features across scales helps the model to better differentiate between road and non-road elements, which contributes to fewer false positives and more true positives. This accuracy in classification and the ability to capture both broad road structures and fine details result in a balanced increase in precision and recall, thereby boosting the F1 score.

Figure 11 presents the results of a qualitative analysis using the CHN6-CUG dataset. The images, arranged from left to right, showcase original test set images, ground truth, and results from Unet, Unet++, DeeplabV3+, D-LinkNet, SWGE-Net, and MSIF-Net. The blue solid lines highlight the most significant parts of the comparison in terms of continuity and boundary completeness of the narrow road extraction results, while the red dashed lines display whether some unlabeled, extremely narrow roads have been successfully extracted. The test images were selected from images containing a large number of narrow roads in Macau, Beijing, Hong Kong, Shanghai, Shen Zhen and Wuhan.

Specifically, in the first, third, and sixth rows, the most significant problem is that the background environment is very complex and contains many different kinds of features, thus posing a challenge to the continuity and completeness of the narrow road extraction results. We can see that SWGE-Net maximally overcomes the complexity of the image background in these three images and obtains the best overall performance, the performance is also largely attributed to the coordinate attention mechanism, which effectively emphasizes important road information and suppresses unnecessary information as much as possible by learning the spatial distribution of feature maps. On the other hand, MSIF-Net successfully extracts the most complete unlabeled extremely narrow roads in the complex background. The most significant problems in the second, fourth and sixth rows are the strong sunlight that may produce high contrast shadows and the occlusion of trees on both sides of the road, which make it difficult to capture the road details. From the results, it can be seen that SWGE-Net demonstrates a performance that significantly outperforms the other networks in terms of continuity and completeness of the road extraction results in the face of shadows and occlusion problems. MSIF-Net

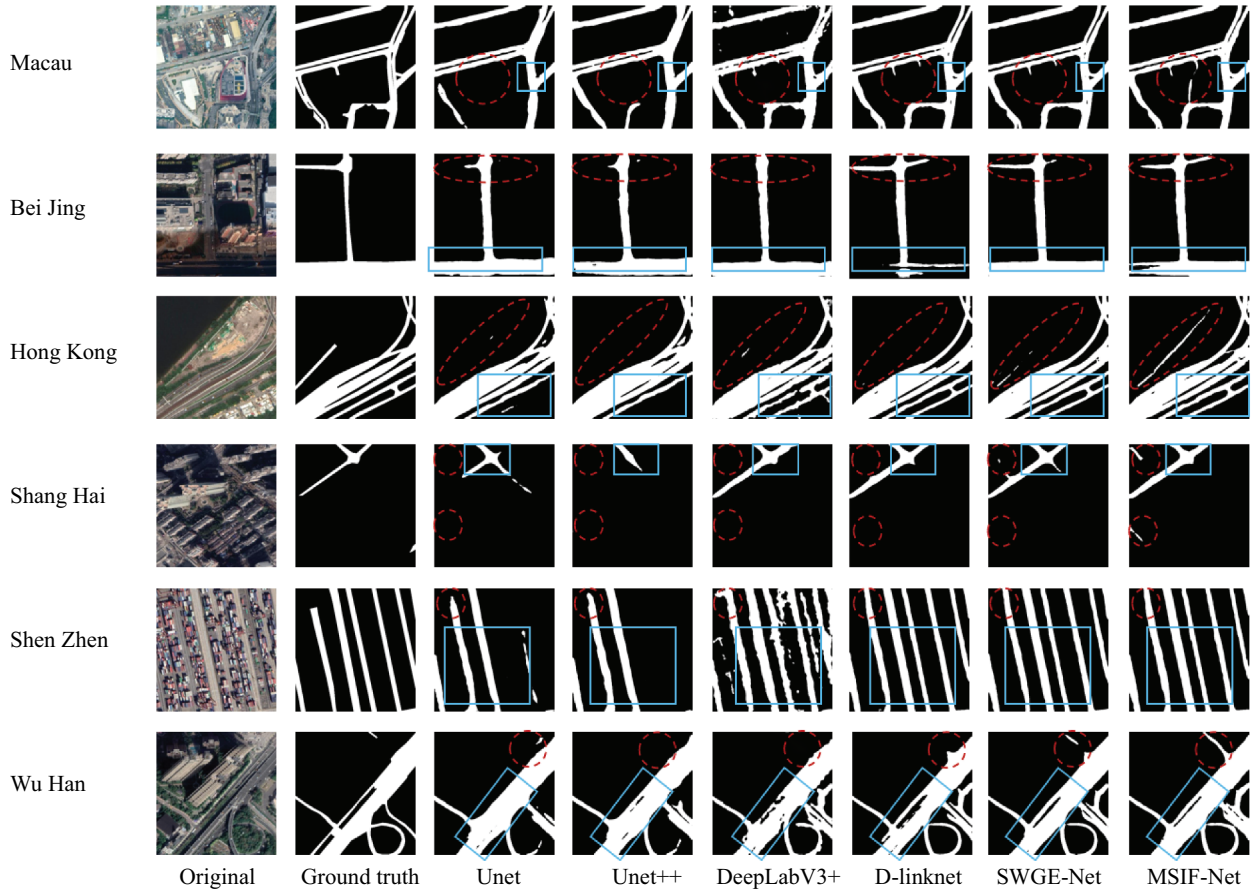

Figure 11. Qualitative evaluation results of different methods on the CHN6-CUG dataset.

successfully extracted the most complete unlabeled extremely narrow roads in the presence of shadows and occlusions, this is due to the powerful cross-scale feature fusion ability of the decoder based on pyramid pooling, which helps the model better construct road features.

4.5. Ablation experiment

To evaluate the effectiveness of the SWGE-Module and MSIF-Module in our proposed methods, we conducted an ablation experiment on the CHN6-CUG dataset. We designed four model configurations: Baseline (DlinkNet34), Baseline + SWGE-Module, Baseline + MSIF-Module, and Baseline + SWGE-Module + MSIF-Module. The performance was assessed using IoU, F1-score, Precision, Recall, and Training Time.

As shown in Table 3, using the SWGE-Module significantly improves all metrics compared to the baseline, with an IoU of 60.67% and an F1-score of 74.25%, demonstrating its effectiveness in enhancing spatial information extraction and detail capture. The MSIF-Module also shows effectiveness with improved performance metrics, achieving an IoU of 59.31% and an F1-score of 72.73%. Additionally, we attempted to combine these two methods into a unified framework. However, the fusion attempts did not result in substantial performance improvements and significantly increased the computational requirements and processing time. The primary challenge lies in the complexity introduced by the combination of the SWGE-Module with the MSIF-Module. Despite the increased model complexity, the fusion did not yield significant performance improvements due to potential redundancy in spatial information capture and risks of over-fitting.

Moreover, Figure 12 shows more qualitative evaluation results of narrow roads extraction to better demonstrate the excellent performance of SWGE-Net and MSIF-Net in narrow roads extraction. The order from left to right is the original image, Ground truth,

Table 3. Ablation experiment on the CHN6-CUG dataset.

Model Name	Backbone	IoU(%)	F1-score(%)	Precision(%)	Recall(%)	Training Time(h)
Baseline	ResNet34	57.56	69.26	72.61	66.21	**19.58**
Baseline+SWGE-Module	ResNet34	**60.67**	**74.25**	**75.69**	**72.86**	27.35
Baseline+MSIF-Module	ResNet34	59.31	72.73	74.71	70.85	41.92
Baseline+SWGE-Module+MSIF-Module	ResNet34	59.05	71.17	74.96	67.74	51.45

In this table, the bold indicates the best result of the objective index in these algorithms.

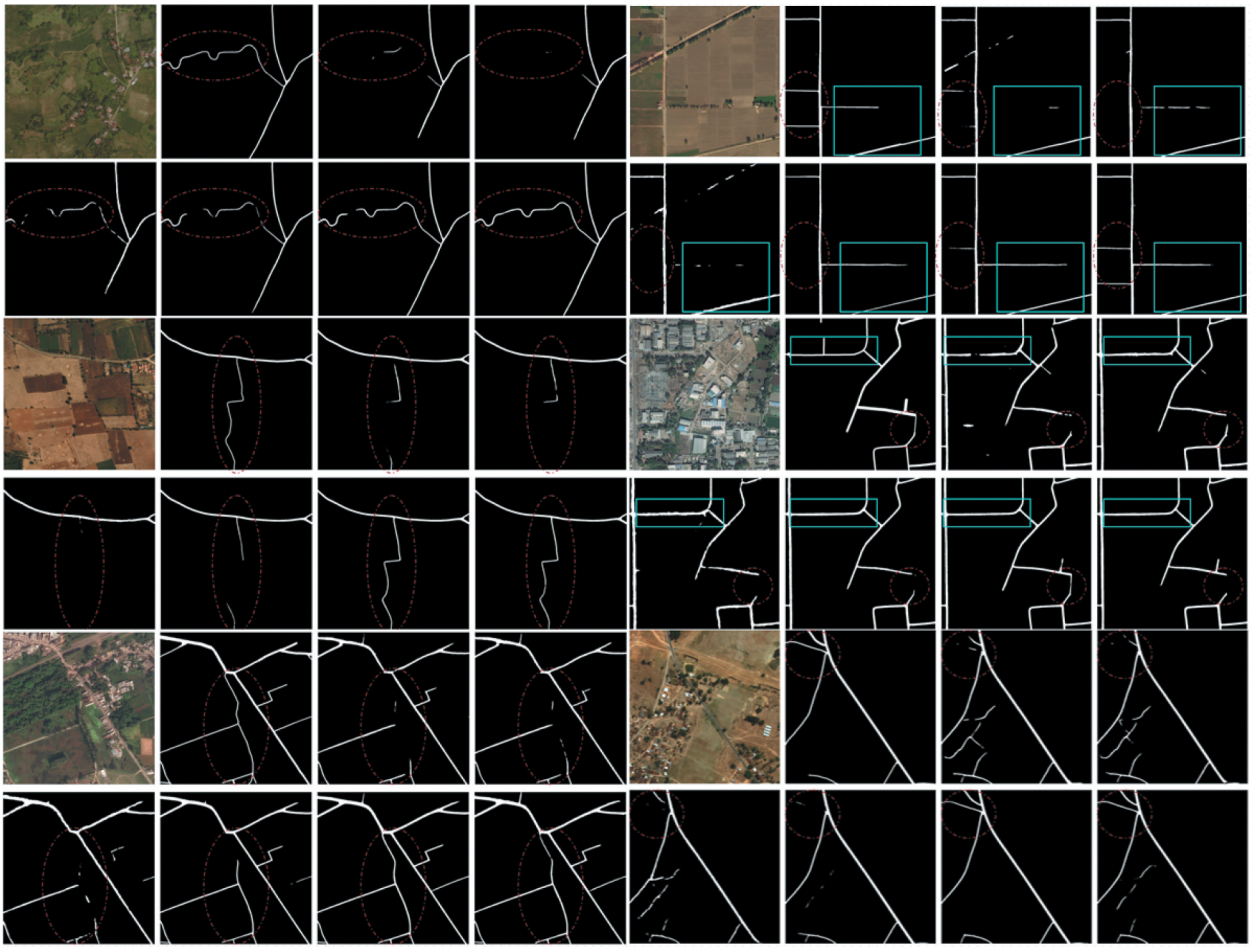

Figure 12. More qualitative evaluation results of narrow roads extraction.

Unet, Unet++, DeeplabV3+, D-linknet, SWGE-Net and MSIF-Net. The solid blue line highlights the parts of the narrow road extraction results that contrast most significantly in terms of continuity and boundary completeness, while the red dashed line shows whether some extremely narrow roads were successfully extracted.

5. Conclusions

In conclusion, the precise extraction of narrow road networks from high-resolution remote sensing images is of critical importance for urban development, navigation systems, and emergency management. This research underscores the significance of addressing the challenge of accurately delineating narrow roads, which are often poorly represented due to their complex morphology and susceptibility to various environmental interferences.

We made significant strides with the introduction of SWGE-Net and MSIF-Net. SWGE-Net's innovative use of a self-weighted global context mechanism resulted in substantial gains in key quantitative metrics, with IoU improvements of 18.51% on the DeepGlobe and 5.40% on the CHN6-CUG datasets over the baseline networks.These improvements are not only evident in qualitative metrics, but also in the qualitative evaluation, where SWGE-Net performs well in maintaining road continuity and boundary integrity under complex environmental conditions. On the other hand, MSIF-Net solves to a certain extent the common problem of unlabeled and extremely narrow roads that are difficult to extract. Through its multi-scale information fusion approach, it demonstrates enhanced capabilities in extracting these challenging road features, surpassing the results of baselines in qualitative evaluations and performing well in extracting tiny roads.

However, this study also identified some areas for improvement. Sometimes, both models struggle to maintain a comparable level of high quality in road extraction results across datasets in different regions, indicating the need to improve their generalization capabilities. Furthermore, the challenge of integrating isolated road segments into an integrated network map remains, especially in heterogeneous terrains and varying conditions such as lighting and seasonal changes.

Looking ahead, our future work will aim to resolve these issues by advancing the generalization capabilities of our networks to function robustly across diverse geographic locales. We plan to introduce sophisticated techniques for seamless road network connectivity and explore the integration of additional data sources such as multi-temporal satellite data to mitigate the effects of environmental factors. The ultimate goal is to push the boundaries of current road extraction technologies, making them more versatile and reliable for real-world applications in various landscapes and conditions.

Disclosure statement

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

Funding

This paper is funded by the Opening Fund of the Key Laboratory of Geological Survey and Evaluation of the Ministry of Education under [grant number GLAB2022ZR02], and the research fund from The Research Institute of Land and Space, Hong Kong Polytechnic University.

Notes on contributors

Zhebin Zhao received his B.S. degrees from Arizona State University, and Hainan University, in 2021, and his M. S. degree from The Hong Kong Polytechnic University, in 2023. He is currently a research assistant in the Department of Land Surveying and Geo-Informatics at The Hong Kong Polytechnic University. His research interests include visual perception, localization and mapping.

Wu Chen received the Ph.D. degree from Newcastle University, in 1992. He is currently a Professor with the Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University. His current research interests include the global navigation satellite system (GNSS) positioning quality evaluation, system integrity, various GNSS applications, seamless positioning, and indoor positioning.

San Jiang received the B.S. degree in remote sensing science and technology and the M.Sc. and Ph.D. degrees in photogrammetry and remote sensing, all from Wuhan University, in 2010, 2012 and 2018, respectively. From 2012 to 2014, he worked as an Assistant Engineer with the Tianjin Institute of Surveying and Mapping. From 2014 to 2015, he was with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing (LIESMARS), Wuhan University, as a Research Assistant. He is currently an Associate Professor with the School of Computer Science, China University of GeoSciences. His research interests include image matching, structure-from-motion-based aerial triangulation, and 3-D reconstruction.

Yaxin Li received the B.S. degree from Wuhan University, in 2013, and the M.S. and Ph.D. degrees from The Hong Kong Polytechnic University, in 2015 and 2020, respectively. His current research interests include indoor positioning and navigation, simultaneous localization and mapping (SLAM), indoor 3-D modeling, semantic segmentation, and auto building information modeling (BIM) generation.

Jingxian Wang received his B.S. degree and M.S. degree from Nanjing University of Aeronautics and Astronautics, in 2015 and 2018. He is currently pursuing the Ph.D. degree with the Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University. His current research interests include multi-sensors fusion, indoor position, pedestrian and vehicle navigation.

ORCID

Zhebin Zhao http://orcid.org/0009-0001-3417-2519
Yaxin Li http://orcid.org/0000-0002-5610-3223
Jingxian Wang http://orcid.org/0000-0001-9106-6344

References

Abdollahi, A., H. R. R. Bakhtiari, and M. P. Nejad. 2018. "Investigation of SVM and Level Set Interactive Methods for Road Extraction from Google Earth Images." *The Journal of the Indian Society of Remote Sensing* 46:423–430. https://doi.org/10.1007/s12524-017-0702-x.

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.

Badrinarayanan, V., A. Kendall, and R. Cipolla. 2017. "Segnet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation." *IEEE Transactions on Pattern Analysis & Machine Intelligence* 39 (12): 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615.

Buslaev, A., S. Seferbekov, V. Iglovikov, and A. Shvets. 2018. "Fully Convolutional Network for Automatic Road Extraction from Satellite Imagery." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, June 18-23, Salt Lake City, UT, USA, 207–210. https://openaccess.thecvf.com/content_cvpr_2018_workshops/w4/html/Buslaev_Fully_Convolutional_Network_CVPR_2018_paper.html.

Chen, L. C., G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. 2014. "Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected Crfs." arXiv preprint arXiv:1412.7062.

Chen, L., Q. Zhu, X. Xie, H. Hu, and H. Zeng. 2018. "Road Extraction from VHR Remote-Sensing Imagery via Object Segmentation Constrained by Gabor Features." *ISPRS International Journal of Geo-Information* 7 (9): 362. https://doi.org/10.3390/ijgi7090362.

Chen, T., D. Jiang, and R. Li. 2022. "Swin Transformers Make Strong Contextual Encoders for VHR Image Road Extraction." *Proceedings of the IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, 3019–3022. IEEE. https://doi.org/10.1109/IGARSS46834.2022.9883628.

Chen, Z., Y. Luo, J. Wang, J. Li, C. Wang, and D. Li. 2023. "DPENet: Dual-Path Extraction Network Based on CNN and Transformer for Accurate Building and Road Extraction." *International Journal of Applied Earth Observation and Geoinformation* 124:103510. https://doi.org/10.1016/j.jag.2023.103510.

Dai, L., G. Zhang, and R. Zhang. 2023. "RADANet: Road Augmented Deformable Attention Network for Road Extraction from Complex High-Resolution Remote-Sensing Images." *IEEE Transactions on Geoscience & Remote Sensing* 61:1–13. https://doi.org/10.1109/TGRS.2023.3237561 .

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." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, June 18-23, Salt Lake City, UT, USA. https://openaccess.thecvf.com/content_cvpr_2018_workshops/w4/html/Demir_DeepGlobe_2018_A_CVPR_2018_paper.html.

Dewangan, D. K., and S. P. Sahu. 2023. "Towards the Design of Vision-Based Intelligent Vehicle System: Methodologies and Challenges." *Evolutionary Intelligence* 16 (3): 759–800. https://doi.org/10.1007/s12065-022-00713-2.

Dewangan, D. K., S. P. Sahu, and K. V. Arya. 2024. "Vision-Sensor Enabled Multi-Layer CNN Scheme and Impact Analysis of Learning Rate Parameter for Speed Bump Detection in Autonomous Vehicle System." *IEEE Sensors* Letters. https://doi.org/10.1109/LSENS.2024.3360095.

Ding, L., Q. Yang, J. Lu, J. Xu, and J. Yu. 2016. "Road Extraction Based on Direction Consistency Segmentation." *Pattern Recognition: 7th Chinese Conference, CCPR 2016, Chengdu, China* edited by T. Tan, X. Li, X. Chen, J. Zhou, J. Yang, and H. Cheng, 131–144, Proceedings, Part I, Singapore. Springer. November 5–7. https://doi.org/10.1007/978-981-10-3002-4_11.

Du, Y., Q. Sheng, W. Zhang, C. Zhu, J. Li, B. Wang, and B. Wang. 2023. "From Local Context-Aware to Non-Local: A Road Extraction Network via Guidance of Multi-Spectral Image." *Isprs Journal of Photogrammetry & Remote Sensing* 203:230–245. https://doi.org/10.1016/j.isprsjprs.2023.07.026.

Gao, X., X. Sun, Y. Zhang, M. Yan, G. Xu, H. Sun, and K. Fu. 2018. "An End-To-End Neural Network for Road Extraction from Remote Sensing Imagery by Multiple Feature Pyramid Network." *Institute of Electrical and Electronics Engineers Access* 6:39401–39414. https://doi.org/10.1109/ACCESS.2018.2856088.

Guo, M. H., T. X. Xu, J. J. Liu, Z. N. Liu, P. T. Jiang, T. J. Mu, S. H. Zhang, R. R. Martin, M. M. Cheng, and S. M. Hu. 2022. "Attention Mechanisms in Computer Vision: A Survey." *Computational Visual Media* 8 (3): 331–368. https://doi.org/10.1007/s41095-022-0271-y.

Hinz, S., and A. Baumgartner. 2000. "Road Extraction in Urban Areas Supported by Context Objects." *International Archives of Photogrammetry and Remote Sensing* 33 (B3/1; PART 3): 405–412. https://www.isprs.org/PROCEEDINGS/XXXIII/congress/part3/405_XXXIII-part3.pdf.

Hou, Q., D. Zhou, and J. Feng. 2021. "Coordinate Attention for Efficient Mobile Network Design." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13713–13722. https://openaccess.thecvf.com/content/CVPR2021/html/Hou_Coordinate_Attention_for_Efficient_Mobile_Network_Design_CVPR_2021_paper.html.

Hou, Y., Z. Liu, T. Zhang, and Y. Li. 2021. "C-UNet: Complement UNet for Remote Sensing Road Extraction." *Sensors* 21 (6): 2153. https://doi.org/10.3390/s21062153.

Huang, X., and L. Zhang. 2009. "Road Centreline Extraction from High-Resolution Imagery Based on Multiscale Structural Features and Support Vector Machines." *International Journal of Remote Sensing* 30 (8): 1977–1987. https://doi.org/10.1080/01431160802546837.

Jie, Y., H. He, K. Xing, A. Yue, W. Tan, C. Yue, C. Jiang, and X. Chen. 2022. "MECA-Net: A MultiScale Feature Encoding and Long-Range Context-Aware Network for Road Extraction from Remote Sensing Images." *Remote Sensing* 14 (21): 5342. https://doi.org/10.3390/rs14215342.

Li, H., K. Qiu, L. Chen, X. Mei, L. Hong, and C. Tao. 2020. "SCAttNet: Semantic Segmentation Network with Spatial

and Channel Attention Mechanism for High-Resolution Remote Sensing Images." *IEEE Geoscience & Remote Sensing Letters* 18 (5): 905–909. https://doi.org/10.1109/LGRS.2020.2988294.

Li, J., Q. Hu, and M. Ai. 2018. "Unsupervised Road Extraction via a Gaussian Mixture Model with Object-Based Features." *International Journal of Remote Sensing* 39 (8): 2421–2440. https://doi.org/10.1080/01431161.2018.1425563.

Li, R., B. Gao, and Q. Xu. 2020. "Gated Auxiliary Edge Detection Task for Road Extraction with Weight-Balanced Loss." *IEEE Geoscience & Remote Sensing Letters* 18 (5): 786–790. https://doi.org/10.1109/LGRS.2020.2985774.

Li, S., C. Liao, Y. Ding, H. Hu, Y. Jia, M. Chen, B. Xu, X. Ge, T. Liu, and D. Wu. 2022. "Cascaded Residual Attention Enhanced Road Extraction from Remote Sensing Images." *ISPRS International Journal of Geo-Information* 11 (1): 9. https://doi.org/10.3390/ijgi11010009.

Lian, R., W. Wang, N. Mustafa, and L. Huang. 2020. "Road Extraction Methods in High-Resolution Remote Sensing Images: A Comprehensive Review." *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing* 13:5489–5507. https://doi.org/10.1109/JSTARS.2020.3023549.

Long, J., E. Shelhamer, and T. Darrell. 2015. "Fully Convolutional Networks for Semantic Segmentation." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3431–3440. https://openaccess.thecvf.com/content_cvpr_2015/html/Long_Fully_Convolutional_Networks_2015_CVPR_paper.html.

Maboudi, M., J. Amini, S. Malihi, and M. Hahn. 2018. "Integrating Fuzzy Object Based Image Analysis and Ant Colony Optimization for Road Extraction from Remotely Sensed Images." *Isprs Journal of Photogrammetry & Remote Sensing* 138:151–163. https://doi.org/10.1016/j.isprsjprs.2017.11.014.

Máttyus, G., W. Luo, and R. Urtasun. 2017. "Deeproadmapper: Extracting Road Topology from Aerial Images." *Proceedings of the IEEE International Conference on Computer Vision*, October 22-29, Venice, Italy, 3438–3446. https://openaccess.thecvf.com/content_iccv_2017/html/Mattyus_DeepRoadMapper_Extracting_Road_ICCV_2017_paper.html.

Miao, Z., W. Shi, P. Gamba, and Z. Li. 2015. "An Object-Based Method for Road Network Extraction in VHR Satellite Images." *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing* 8 (10): 4853–4862. https://doi.org/10.1109/JSTARS.2015.2443552 .

Mosinska, A., P. Marquez-Neila, M. Koziński, and P. Fua. 2018. "Beyond the Pixel-Wise Loss for Topology-Aware Delineation." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3136–3145. https://openaccess.thecvf.com/content_cvpr_2018/html/Mosinska_Beyond_the_Pixel-Wise_CVPR_2018_paper.html.

Ronneberger, O., P. Fischer, and T. Brox. 2015. "U-Net: Convolutional Networks for Biomedical Image Segmentation." *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference*, 234–241, Proceedings, Part III, Munich, Germany, October 5–9. https://doi.org/10.1007/978-3-319-24574-4_28.

Shamsolmoali, P., M. Zareapoor, H. Zhou, R. Wang, and J. Yang. 2020. "Road Segmentation for Remote Sensing Images Using Adversarial Spatial Pyramid Networks." *IEEE Transactions on Geoscience & Remote Sensing* 59 (6): 4673–4688. https://doi.org/10.1109/TGRS.2020.3016086.

Tao, C., J. Qi, Y. Li, H. Wang, and H. Li. 2019. "Spatial Information Inference Net: Road Extraction Using Road-Specific Contextual Information." *Isprs Journal of Photogrammetry & Remote Sensing* 158:155–166. https://doi.org/10.1016/j.isprsjprs.2019.10.001.

Tao, J., Z. Chen, Z. Sun, H. Guo, B. Leng, Z. Yu, Y. Wang, Z. He, X. Lei, and J. Yang. 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.

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 & Transportation Engineering* 3 (3): 271–282. https://doi.org/10.1016/j.jtte.2016.05.005.

Wei, Y., and S. Ji. 2021. "Scribble-Based Weakly Supervised Deep Learning for Road Surface Extraction from Remote Sensing Images." *IEEE Transactions on Geoscience & Remote Sensing* 60:1–12. https://doi.org/10.1109/TGRS.2021.3061213.

Xu, Y., Z. Xie, Y. Feng, and Z. Chen. 2018. "Road Extraction from High-Resolution Remote Sensing Imagery Using Deep Learning." *Remote Sensing* 10 (9): 1461. https://doi.org/10.3390/rs10091461.

Zhang, R., W. Zhu, Y. Li, T. Song, Z. Li, W. Yang, L. Yang, T. Zhou, and X. Xu. 2023. "D-Fusionnet: Road Extraction from Remote Sensing Images Using Dilated Convolutional Block." *GIScience & Remote Sensing* 60 (1): 2270806. https://doi.org/10.1080/15481603.2023.2270806.

Zhang, Z., Q. Liu, and Y. Wang. 2018. "Road Extraction by Deep Residual U-Net." *IEEE Geoscience & 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. https://doi.org/10.3390/app12073511.

Zhang, Z., X. Sun, and Y. Liu. 2022. "GMR-Net: Road-Extraction Network Based on Fusion of Local and Global Information." *Remote Sensing* 14 (21): 5476. https://doi.org/10.3390/rs14215476.

Zhao, H., J. Shi, X. Qi, X. Wang, and J. Jia. 2017. "Pyramid Scene Parsing Network." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2881–2890. https://openaccess.thecvf.com/content_cvpr_2017/html/Zhao_Pyramid_Scene_Parsing_CVPR_2017_paper.html.

Zhao, L., L. Ye, M. Zhang, H. Jiang, Z. Yang, and M. Yang. 2023. "DPSDA-Net: Dual-Path Convolutional Neural Network with Strip Dilated Attention Module for Road Extraction from High-Resolution Remote Sensing Images." *Remote Sensing* 15 (15): 3741. https://doi.org/10.3390/rs15153741.

Zhou, L., C. Zhang, and M. Wu. 2018. "D-Linknet: LinkNet with Pretrained Encoder and Dilated Convolution for High Resolution Satellite Imagery Road Extraction." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 182–186. https://openaccess.thecvf.com/content_cvpr_2018_workshops/w4/html/Zhou_D-LinkNet_LinkNet_With_CVPR_2018_paper.html.

Zhu, Q., Y. Zhang, L. Wang, Y. Zhong, Q. Guan, X. Lu, L. Zhang, and D. Li. 2021. "A Global Context-Aware and Batch-Independent Network for Road Extraction from VHR Satellite Imagery." *Isprs Journal of Photogrammetry & Remote Sensing* 175:353–365. https://doi.org/10.1016/j.isprsjprs.2021.03.016.