frontiers | Frontiers in Earth Science

TYPE Original Research
PUBLISHED 28 December 2023
DOI 10.3389/feart.2023.1301281

OPEN ACCESS

EDITED BY
Alex Hay-Man Ng,
Guangdong University of Technology, China

REVIEWED BY
Chaoguang Men,
Harbin Engineering University, China
Fu Ren,
Wuhan University, China
Ruisheng Wang,
University of Calgary, Canada

*CORRESPONDENCE
Chao Ma,
✉ 594857325@qq.com

RECEIVED 24 September 2023
ACCEPTED 31 October 2023
PUBLISHED 28 December 2023

CITATION
Xiong S, Ma C, Yang G, Song Y, Liang S and Feng J (2023), Semantic segmentation of remote sensing imagery for road extraction via joint angle prediction: comparisons to deep learning. *Front. Earth Sci.* 11:1301281. doi: 10.3389/feart.2023.1301281

Semantic segmentation of remote sensing imagery for road extraction via joint angle prediction: comparisons to deep learning

Shun Xiong[1], Chao Ma[1]*, Guang Yang[2], Yaodong Song[3], Shuaizhe Liang[4] and Jing Feng[5]

[1]State Key Laboratory of Geo-Information Engineering, Xi'an, China, [2]The First Geoinformation Mapping Institute of MNR, Xi'an, China, [3]Institute of Surveying and Mapping Standardization, MNR, Xi'an, China, [4]School of Computer Science, Beijing University of Technology, Beijing, China, [5]Star Map Press, Beijing, China

Accurate road network information is required to study and analyze the relationship between land usage type and land subsidence, and road extraction from remote sensing images is an important data source for updating road networks. This task has been considered a significant semantic segmentation problem, given the many road extraction methods developed for remote sensing images in recent years. Although impressive results have been achieved by classifying each pixel in the remote sensing image using a semantic segmentation network, traditional semantic segmentation methods often lack clear constraints of road features. Consequently, the geometric features of the results might deviate from actual roads, leading to issues like road fractures, rough edges, inconsistent road widths, and more, which hinder their effectiveness in road updates. This paper proposes a novel road semantic segmentation algorithm for remote sensing images based on the joint road angle prediction. By incorporating the angle prediction module and the angle feature fusion module, constraints are added to the angle features of the road. Through the angle prediction and angle feature fusion, the information contained in the remote sensing images can be better utilized. The experimental results show that the proposed method outperforms existing semantic segmentation methods in both quantitative evaluation and visual effects. Furthermore, the extracted roads were consecutive with distinct edges, making them more suitable for mapping road updates.

KEYWORDS

angle prediction, semantic segmentation, road extraction, remote sensing image, map cartography

1 Introduction

In recent years, land subsidence has led to an increase in road collapses and related accidents, indicating the need to pay close attention to the relationship between roads and subsidence. Land subsidence, also known as ground sinking, is a localized downward movement or geotechnical phenomenon resulting in a lower elevation of the crust surface. This can be influenced by human activities, particularly due to the consolidation

and compression of underground loose stratum, and it often correlates with land use patterns (Xin et al., 2022; Zainuri et al., 2022). The road network is an important indicator of urban development and represents a significant type of urban land use. Therefore, accurately determining the extent of the road network is crucial when investigating the relationship between land use and land subsidence.

In order to accurately extract the road information, many studies have used remote sensing images as an important data source for road extraction in recent years (Yan and Gulimila, 2023). Road extraction from remote sensing images aims to automatically detect and extract road information from remote sensing images. Although many methods have been proposed, the existing methods still have some problems in generating geometric features of roads (Fei and Man, 2021). For example, the semantic segmentation method is usually used for road extraction of remote sensing images, but it lacks clear road feature constraints, resulting in road fractures, rough edges, and inconsistent road widths (Deng et al., 2020).

The road extraction method of remote sensing images based on deep learning has developed rapidly in recent years. Among them, the extraction method of road regions based on semantic segmentation requires remote sensing images to be trained with data corresponding to road masks one by one, while the extraction method of road centerline based on vector maps requires remote sensing images to be trained with corresponding road vector information. Most network map operators provide network map information corresponding to remote sensing images, and network maps usually use different colors with discrimination to mark different types of features on the map. Therefore, the clustering method based on color value or gray level can be used to extract road mask information from the network map. In addition, there are many public datasets containing remote sensing images and road masks as comparison benchmarks in the academic world, such as DeepGlobe (DEEPGLOBE, 2023), SpaceNet (SpaceNetChallenge, 2023), etc., but the road vector information is relatively difficult to obtain. The most common and open vector road information source is the Open Street Map service at present, and the road extraction method based on road vector information uses the vector data set self-extracted from Open Street Map (Robert et al., 2013). However, the quality of vector data extracted from this map service may be inconsistent with that of ordinary network maps in some regions. Given the challenges involved in obtaining road vector information, the aim is to make the proposed method more versatile. This adaptability will further aid in future map generation tasks based on remote sensing images. This paper mainly focuses on the road area extraction method based on semantic segmentation, and studies it from the perspective of considering the task as a binary semantic segmentation task.

Road extraction is one of the most important applications of remote sensing images, and it is also a significant data source for mapping road updates. Early extraction methods usually relied on the constructed recognition factors including texture features, geometric features, and topological features of roads, and the methods of edge extraction and template matching were used to extract roads, after the preprocessing of remote sensing images, which already acquired a certain achievements. At present, there are various kinds of classifications for these traditional road extraction methods in the academic field, the traditional road extraction methods can be divided into three categories: template matching methods, knowledge-based methods, and object-based analysis methods, according to the basic principle (Cheng and Han, 2016; Dai et al., 2020). Firstly, the method based on template matching (Bajcsy and Tavakoli, 1976; Vosselman and Knecht, 1995; Rathinam et al., 2008) is a relatively mature method in the early road extraction methods, which can effectively combine the radiation characteristics and geometric characteristics of a road in order to assess, can achieve the human-computer interaction by setting the seed point or template initial contour, has a certain ability of error correction, and has already been widely used. Secondly, the object-based analysis method (Drăguț et al., 2014; Sheng et al., 2015) relies on the fixed rules of artificial or semi-artificial priors, the detection effect is still susceptible to complex real situations when facing the high-resolution remote sensing images. Finally, the knowledge-based method (Willrich, 2002; Herumurti et al., 2013) requires the use of some auxiliary knowledge to carry out or assist in road extraction, but this knowledge is relatively complicated, and it is usually necessary to manually set *a priori* conditions to use it. Therefore, the application of this method has low usage.

With the prosperity and wide application of deep learning technology in the field of computer vision, various extraction algorithms based on convolutional neural networks have emerged, and have achieved better extraction results than traditional methods. In particular, algorithms represented by semantic segmentation have gradually received extensive attention and application due to their advantages in extracting road details. This method regards the road extraction problem as a binary semantic segmentation problem of remote sensing images, and uses a semantic segmentation model with a deep convolutional neural network as the core to divide each pixel in remote sensing images into road or background categories. While the network structure designed for natural image semantic segmentation is valuable, it encounters challenges when applied to road extraction. Issues such as road fractures and blurred road boundaries persist in the extraction results (Saito et al., 2016; Zhong et al., 2016). Typically, the semantic segmentation network model lacks constraints to address these specific problems. Therefore, the rule template or context knowledge of the traditional method is combined as a constraint condition, further improving the quality of road extraction (Cheng et al., 2017; Wei et al., 2017; Li et al., 2021; Wan et al., 2021; Zhu et al., 2021). In addition, there is a kind of road extraction method that uses the topological structure of road as a constraint condition. The neural network is used to predict the iterative target point, and the graph is connected after multiple iterations and used as the predicted road network result (Bastani et al., 2018; Tan et al., 2020). This kind of method has greater advantages in predicting the coherence of a road but has higher requirements for training data and prediction time. Although the deep learning method has the advantages of strong generalization and a high degree of automation, it also has some problems: it requires many accurate data sets, and the training time of deep learning is too slow. At present, the model method is still in the laboratory research stage, and a wide range of large-scale road extraction based on deep learning has not yet been carried out.

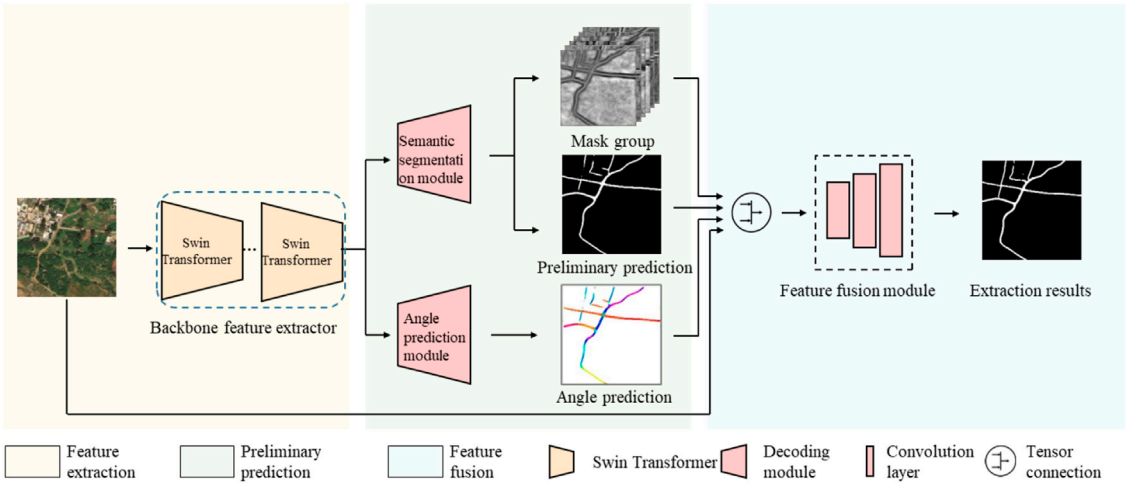

FIGURE 1
Overall structure of a semantic segmentation network.

To summarize, when the current semantic segmentation method is used to extract roads from remote sensing images, the issues of road fractures, rough edges, and inconsistent road width are still common in the extraction results. In order to solve these problems and improve the accuracy and effect of road extraction from remote sensing images, this study considers the prediction and fusion of angle information from the geometric characteristics of road inclination angle with a certain amount of stability and uses angle information to assist road extraction and improve its effect. Thus, a road extraction method for remote sensing images based on semantic segmentation and angle prediction is consequently proposed in this paper. Based on the semantic segmentation method, the angle prediction module and the angle feature fusion module were added. By adding the constraints to the angle features of the road, prediction and feature fusion were performed to make better use of the information contained in the remote sensing image. Compared with the existing methods, the final extracted roads generated by this method were consecutive and had clear boundaries, which can be used for subsequent map road network updates.

In order to solve the above problems, a road extraction method of remote sensing images combining semantic segmentation and angle prediction is proposed in this paper. The proposed method enhances the traditional semantic segmentation approach by integrating both an angle prediction module and an angle feature fusion module at its core. By using the angle information of the road as a constraint, prediction and feature fusion are performed, optimizing the utilization of information from the remote sensing image. Specifically, the angle prediction module is mainly used to predict the inclination angle of the road in the remote sensing image, while the angle feature fusion module is used to fuse the predicted angle features with the semantic segmentation results, to obtain more accurate road extraction results. The experimental results show that the combined method is superior to the existing semantic segmentation method and road extraction method in quantitative evaluation and visual effect. By introducing angle information as a constraint, this method can effectively solve the problems of road fractures, rough edges, and inconsistent road widths, so as to produce more accurate and continuous road extraction results. These results have important application value in the fields of urban planning, traffic management, and navigation systems.

2 Methodology

2.1 Overall network structure

At present, due to the need for remote sensing images and road masks requiring one-to-one correspondence for training, the road extraction methods based on semantic segmentation usually use deep convolutional neural networks as the backbone network, which has maintained excellent performance in semantic segmentation tasks over the past few years. Recently, a model structure called Transformer that has achieved better results in the field of natural language processing has been migrated to computer vision tasks, and has achieved good results in mainstream computer vision problems such as target detection, semantic segmentation, instance segmentation, etc. (Vaswani et al., 2017). The network in this paper is designed in a mode of multi-segment Transformer combination, and the overall structure of the network is shown in Figure 1.

In the network backbone part, referring to the Swin Transformer method (Liu et al., 2021), the structural segments of multiple Swin Transformers are firstly stacked as the network backbone feature extractor, to extract the features of the image. The features are then transmitted to the semantic segmentation module and the angle prediction module, respectively. The goal of the semantic segmentation module is to make a preliminary prediction of road region, and the angle prediction module predicts the road angle of possible road region. Based on the unique geometric characteristics of the road, the inclination of the road can be used as important information to judge whether the road prediction is reasonable. Finally, the features of the two modules are fused to complete the task of road extraction.

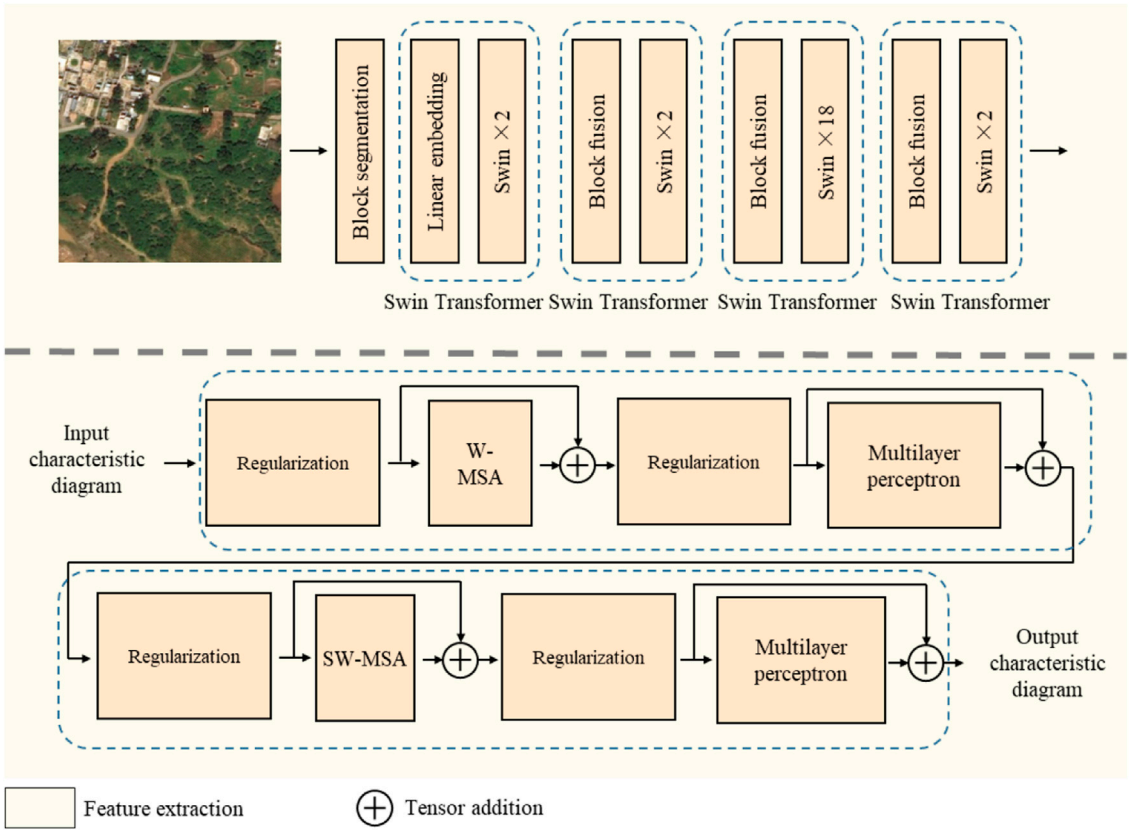

FIGURE 2
Structure of backbone feature extractor.

2.2 Backbone feature extractor

The structure of a backbone feature extractor is shown in Figure 2. Firstly, the input H × W × 3 RGB remote sensing image is divided into the form of N × p2 × 3 by using the block cutting layer, that is, it is divided into N blocks, and the size of each block is p2 × 3. Secondly, the p2 × 3 dimensional tensor of each block is projected to the vector mapping of any dimension C by the linear embedding layer, where the linear embedding layer is essentially a fully connected layer. The significance of this step is to map an RGB small block of each p2 × 3 to a linear vector for the use of the subsequent Transformer structure. Subsequently, these vectors are input into the Swin block of the self-attention mechanism. The lower part of Figure 2 shows the internal structure of two consecutive adjacent Swin blocks. These vectors are input into the multi-head self-attention module W-MSA (Windowing Multi-head Self Attention) according to the windows or multi-head self-attention module SW-MSA (Shifted Windowing Multi-head Self Attention) based on the shifted windows after regularization.

The standard Transformer structure uses a global self-attention module, and in the image task, the image features have a large vector dimension, so the global self-attention module has an amazing computational cost. The multi-head self-attention module is based on windows taking the block cut by layer as the unit, the self-attention is calculated inside each block, which significantly improves the computational efficiency of the model, but limits the information exchange cross-windows. Therefore, the multi-head self-attention module based on the windows and the multi-head self-attention module based on the shift windows are alternately used in each continuous two adjacent Swin blocks. In the multi-head self-attention module based on the shift windows, the 1/2 size of the window is shifted in the horizontal and vertical direction, respectively, the self-attention is calculated, and then the reverse shift is performed, to complete the information exchange between the windows. After the self-attention module, the residual connection is performed in the model, and continues to use the regularization layer and the multi-layer perceptron layer. Multilayer perceptron is the basic structure of deep learning, which takes the full connection as the basic principle. In the following Swin structure segments of the main feature extractor, the segment fusion layer instead of the linear embedding layer is performed in the model, which conducts the down-samples for the current feature map at the front of each Swin structure segment.

2.3 Decoding module

After using the main feature extractor to extract the feature map from the remote sensing image, the semantic segmentation module and the angle prediction module are used to decode the feature map, and the road preliminary extraction and angle prediction are carried out, respectively. The structure is shown in Figure 3. In order to

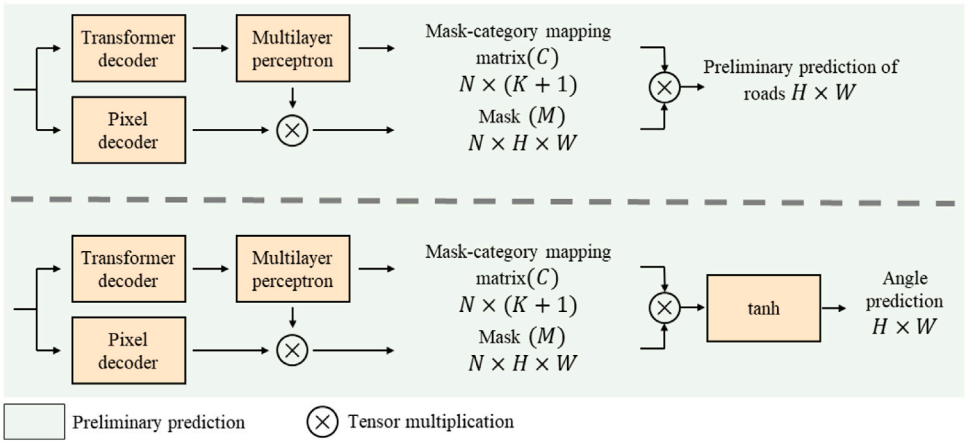

FIGURE 3
Structure of decoding module.

improve the prediction effect, the pixel decoder and the standard Transformer decoder are used in parallel to predict two sets of values: a set of mask M and a set of mask-to-category mapping matrix C (Cheng et al., 2021a). The mask M contains N masks, and the size of each mask is H × W, which is consistent with the input image. The size of mapping matrix C is N × (K+1), which connects N masks with the distribution of K categories to be predicted by the matrix multiplication with the mask group M. The pixel decoder uses traditional convolution up-sampling, while the Transformer decoder (Carion et al., 2020) uses multiple attention layers for up-sampling. In the angle prediction module, an additional tanh layer is used to constrain the output to − 1~1, which is then multiplied by π/2 to map to the angle space. The semantic segmentation module and the angle prediction module will obtain preliminary road region prediction and road angle prediction, respectively, which will be compared with the corresponding true values to calculate the loss function.

2.4 Feature fusion module

After obtaining the preliminary road region prediction and road angle prediction, the road region prediction features and the road angle features are fused by the feature fusion module, to further improve the accuracy of road prediction. The feature fusion module consists of multiple convolutional layers, and the preliminary road prediction results, angle prediction results, and multi-channel feature maps connected by the mask group M are up-sampled by these convolutional layers, to obtain the final road prediction results.

2.5 Loss function design

In order to effectively train the model, three loss functions are used at different modules, that is the loss function of preliminary road prediction, the loss function of angle prediction, and the final loss function of road prediction.

(1) Loss function of preliminary road prediction. In the semantic segmentation, the preliminary road prediction task is decomposed into two sub-tasks, which predict the mask group M and the mapping matrix C, so it is necessary to set up the loss functions, respectively. The mask group M uses a combination of focus loss and dice loss; The mapping matrix C uses the cross-entropy classification loss. The loss function $L_{mask-cls}$ that constrains the preliminary road prediction results can be expressed as:

$$\mathcal{L}_{mask-cls} = \mathcal{L}_{CE}(c, c^{gt}) + \mathcal{L}_{foc}(m, m^{gt}) + \mathcal{L}_{dice}(m, m^{gt}) \quad (1)$$

Among them, $\mathcal{L}_{CE}$, $\mathcal{L}_{foc}$, and $\mathcal{L}_{dice}$ are loss functions of cross entropy classification, focus loss function, and dice loss function, respectively; c and c^{gt} represent the predicted results and true values of mapping matrix C, respectively; m and m^{gt} represent the predicted results and true values of mask group M, respectively.

(2) Loss function of angle prediction. Although the structure of angle prediction module is similar to the semantic segmentation module, the prediction target is different. The module predicts the angle value corresponding to each road pixel, which is represented by the value between −π/2 ~ π/2. Therefore, the loss function of constraint angle prediction results is designed, which can be shown in formula 2.

$$\mathcal{L}_{theta} = \sum_{(x,y)|r^{gt}_{(x,y)}=1} min\left(\left|t_{(x,y)} - t^{gt}_{(x,y)}\right|, \pi - \left|t_{(x,y)} - t^{gt}_{(x,y)}\right|\right) \quad (2)$$

Where r^{gt} represents the true value of the road region, t and t^{gt} represent the predicted results and the true value of the road angle matrix, respectively.

(3) Loss function of final road prediction. The multiple convolutional layers are adopted by the feature fusion module for up-resample, and the results can be regarded as pixel-level semantic segmentation results. Therefore, the loss function of the pixel classification type is used to constrain the

final prediction results, that is, the focus loss and lovász loss constraint model are used, which can be shown in formula 3.

$$\mathcal{L}_{ret} = \mathcal{L}_{foc}\left(r, r^{gt}\right) + \mathcal{L}_{lov}\left(r, r^{gt}\right) \tag{3}$$

3 Experiment and analysis

3.1 Experimental data

The experimental data is from the Deep Globe public dataset, which is widely used in the road extraction task of remote sensing images (Demir et al., 2018). The Deep Globe dataset is derived from the CVPR Deep Globe 2018 road extraction challenge, which consists of 6,226 pairs of training images, 1,243 pairs of validation images, and 1,101 pairs of experimental images. The image size is 1,024 × 1,024 pixels and the spatial resolution is 0.5 m. In order to facilitate the comparison with other road extraction methods, the original Deep Globe dataset is processed according to the reference (Singh et al., 2018). The 6,226 pairs of training images in the original dataset are re-split and divided into multiple 512 × 512 blocks at a 256-pixel interval.

3.2 Evaluation index of experimental results

The accuracy of the road extraction results was evaluated by an F1 score and IoU score. The F1 score is an indicator used to measure the accuracy of binary classification in statistics, which is calculated based on accuracy (P) and recall ratio (R), and the formula can be expressed as follows:

$$F1 = \frac{2 \times P \times R}{P + R}, P = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN} \tag{4}$$

TP, FP, TN, and FN represent the number of true positive, false positive, true negative, and false negative pixels, respectively.

IoU is a commonly used evaluation index in semantic segmentation, which is the ratio of intersection acreage to the union acreage between the real region and the predicted region corresponding to a semantic category, and the calculation formula can be expressed as:

$$IoU = \frac{TP}{TP + FP + FN} \tag{5}$$

3.3 Experimental results

The processed Deep Globe data are input into the constructed model to complete the model training and testing. In order to illustrate the advancement of the proposed algorithm, it is compared with the traditional fully connected neural network method, the existing Transformer model method, the semantic segmentation method, and the road extraction method popular in recent years, respectively, and the results are shown in Figure 4. Quantitative evaluation is performed according to the result evaluation index in Section 4.2, and the results are shown in Table 1.

The comparative semantic segmentation methods include DeepLabv3 + (Chen et al., 2018), DeepUNet (Li et al., 2018), and BRRNet (Shao et al., 2020), all of which are based on fully convolutional neural networks. Additionally, methods based on Transformer structure, such as Swin Transformer (Liu et al., 2021), MaskFormer (Cheng et al., 2021b), and Mask2Former (Cheng et al., 2021a) are also included. Furthermore, recent high-performing road extraction methods like CoANet (Mei et al., 2021), LinkNet50 + GA (Lu et al., 2020), GAMSNet (Lu et al., 2020), and GCB-Net (Zhu et al., 2021) have also been selected for comparison. Among them, for the method of CoANet, its public code was used to train the model and evaluate the results under the same batch processing size and similar total number of iterations as other comparative experiments. For the methods of LinkNet50 + GA, GAMSNet, and GCB-Net, the results scores provided in their respective papers from the Deep Globe public dataset were used for comparison (Lu et al., 2020; Lu et al., 2020; Zhu et al., 2021) as their codes have not been made publicly available. However, the results images are not compared in this work. The method based on the Transformer structure is superior to the method based on the full convolutional neural network, indicating that the Transformer structure has a very good advantage in extracting features. After adding the angle prediction module, the road contour extracted by our method is more intuitive than other methods based on Transformer structure. There are further comparative experiments in the analysis of experimental results. It can be seen from Figure 4 that the road contour extracted by the method proposed in this paper is more intuitive, and the quantitative evaluation results in Table 1 also show the superiority of this method in this paper.

3.4 Experimental analysis

In this section, several aspects of proposed method are discussed.

3.4.1 Selection of fusion features

In order to study the influence of features selected for fusion in the feature fusion layer on the final result, the features used for fusion are adjusted, and the training results are observed. In the main experiment, the original remote sensing image, the preliminary road prediction result, the road angle prediction result, and the mask group M generated in the preliminary road prediction process are input into the feature fusion module, to fuse and predict the final road prediction result. In this experiment, each feature for fusion is removed and the model is trained under the same other settings, to verify the utility of each feature. The test results are shown in Table 2, and it can be seen that each feature used for fusion has a positive impact on the final road prediction results.

3.4.2 Comparison of parameters quantum

In this experiment, the parameter quantity is used by the network model and comparison method as the index to compare the model complexity. The parameter quantity and results of different models trained on the DeepGlobe dataset are shown in Table 3. The method proposed in this paper in the case of optimal results, is lower than the Swin Transformer method, and slightly higher than the MaskFormer and Mask2Former methods in the parameter quantity. This shows that the model performance is not only improved by stacking more parameter quantum for the method proposed by this paper, but also can further optimize the perspective of parameter quantity.

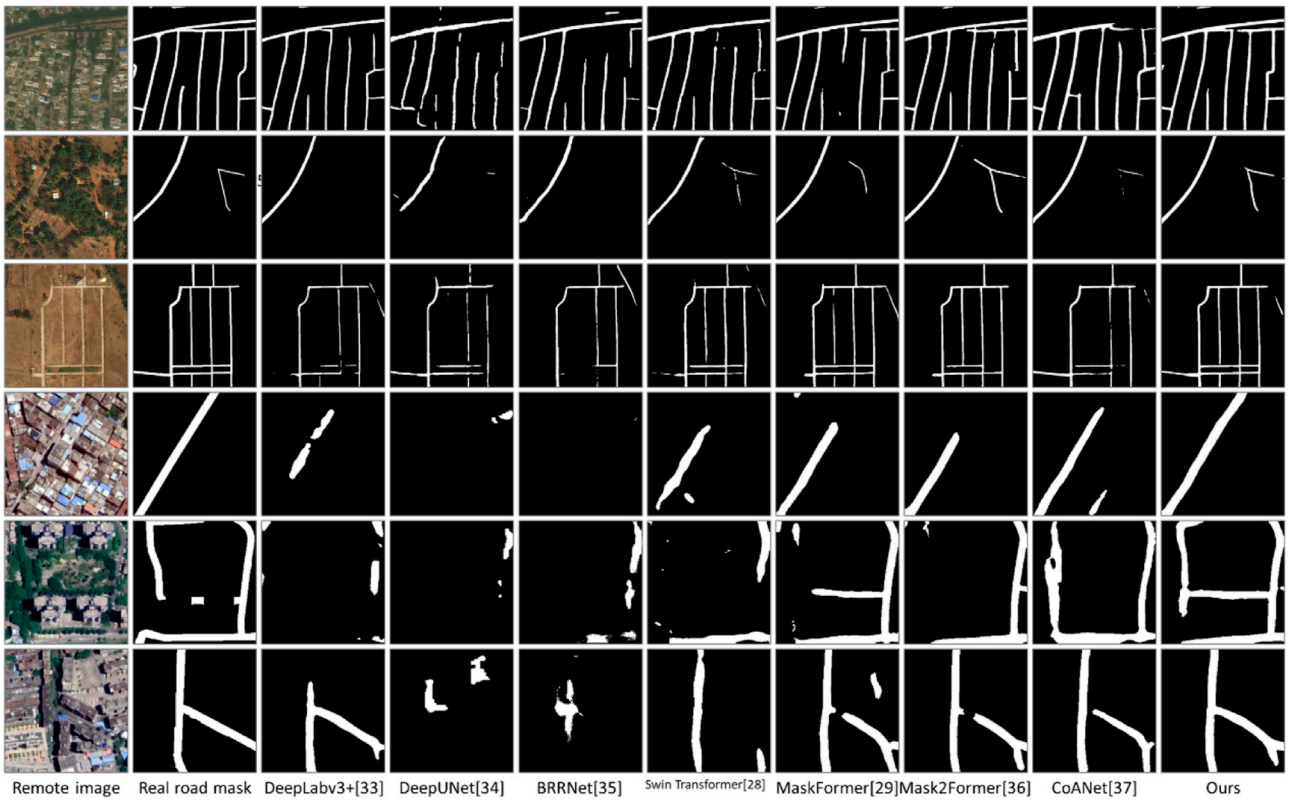

FIGURE4
Experimental result.

TABLE 1 Comparison of different methods on DeepGlobe public dataset.

Model	IoU score (road category)	F1 score (road category)
DeepLabv3+	0.6612	0.7960
DeepUNet	0.4608	0.6309
BRRNet	0.5708	0.7267
Swin Transformer	0.6378	0.7788
MaskFormer	0.7110	0.8311
Mask2Former	0.7077	0.8288
CoANet	0.6459	0.7848
LinkNet50+GA	0.6821	0.8110
GAMSNet	0.6945	0.8197
GCB-Net	0.7080	0.8154
Ours	**0.7132**	**0.8323**

Bold value means the best result of the experiment.

3.4.3 Predicting the number of masks in the mask group

In the semantic segmentation module and angle prediction module, the pixel decoder and standard Transformer decoder are used in parallel to predict the mask group M and the mask-to-category mapping matrix C, respectively, and then both are subjected to matrix multiplication to obtain the scheme to be predicted. The mask group consists of N masks, and the size of each mask is consistent with the remote sensing image, while the value of N has no clear relationship with the task itself, which is a manually set hyper-parameter. For the natural image semantic segmentation task in the reference (Cheng et al., 2021b), the

TABLE 2 Comparison of different feature fusion selection methods.

Model	IoU score (road category)	F1 score (road category)
Original remote sensing image + preliminary road prediction result + road angle prediction result + mask group (main experiment selection)	0.7152	0.8340
Preliminary road prediction results + road angle prediction results + mask group	0.6486	0.7868
Original remote sensing image + road angle prediction result + mask group	0.7128	0.8323
Original remote sensing image + preliminary road prediction result + mask group	0.7142	0.8333
Original remote sensing image + preliminary road prediction result + road angle prediction result	0.7147	0.8336

TABLE 3 Comparison of different methods on DeepGlobe public dataset.

Model	Number of parameter (million)	IoU score (road category)	F1 score (road category)
DeepLabv3+	54.71	0.6612	0.7960
DeepUNet	2.44	0.4608	0.6309
BRRNet	17.34	0.5708	0.7267
Swin Transformer	121.17	0.6378	0.7788
MaskFormer	101.79	0.7110	0.8311
Mask2Former	106.92	0.7077	0.8288
CoANet	59.15	0.6459	0.7848
Ours	**117.12**	**0.7132**	**0.8323**

Bold value means the best result of the experiment.

TABLE 4 Comparison of the number of different masks in the predicted mask group.

Number of masks N	IoU score (road category)	F1 score (road category)
10	0.7129	0.8324
20	0.7137	0.8329
50	0.7130	0.8325
100(the main experiment selection)	**0.7152**	**0.8340**
200	0.7147	0.8336
500	0.7146	0.8336

Bold value means the best result of the experiment.

value of N set to 100 is a better choice; however, the datasets faced by the model in this paper have 150–847 semantic categories, with an average of 6.6–9.1 semantic categories peer image, which is far from the requirements of road prediction and angle prediction tasks in this paper. Therefore, the experiment is re-organized in this paper to test the influence of different N values on road prediction results. It can be seen from Table 4 that the N set to 100 is still a better choice, because that reducing N value and increasing N value will bring different degrees of effect decline.

4 Discussion

The research of remote sensing images in road semantic segmentation has made remarkable progress. By using the high resolution and wide coverage of remote sensing images, the road information can be effectively extracted by researchers and achieve accurate segmentation of roads. The current research mainly focuses on two aspects: feature extraction and classification algorithms. In terms of feature extraction, a variety of different methods have been used by researchers, including color, texture, shape and spatial information. These features can effectively capture the different attributes of the road and provide strong support to achieve accurate segmentation. In terms of classification algorithms, many different methods have been adopted by researchers, including traditional machine learning algorithms and deep learning algorithms. Traditional machine learning algorithms such as support vector machine and random forest have achieved some success in road semantic segmentation, but their performance is relatively low due to their limitations on feature expression. On the other hand, deep

learning algorithms such as convolutional neural networks and recurrent neural networks can better capture the complex features of roads, thereby achieving more accurate segmentation.

The contribution of this paper includes the proposition of a road semantic segmentation extraction method for remote sensing images considering angle prediction. This method uses the stacked Swin Transformer as the feature extraction model, the road angle prediction module is innovatively added to the model, using the prior information of the road angle as a constraint condition, the road angle prediction information is fused with the preliminary prediction of road features, and finally the road segmentation result is obtained.

By comparing with other methods on different remote sensing datasets, results show that both the IoU score and the F1 score are relatively high, which indicates that the effectiveness and superiority of the proposed method. Secondly, the paper proves the superiority of using road angle information as the constraint condition of road extraction related tasks, which can also provide a reference for other related semantic segmentation tasks or other road extraction tasks.

It is noteworthy that if the network model and the comparison method are compared with the parameter quantity as the index, the complexity of the model is comparable. The method in this paper in the case of optimal effect is lower than the Swin Transformer method, and slightly higher than the MaskFormer and Mask2Former methods for the parameter quantity. This shows that the method does not only improve the performance of the model by stacking more parameters, but also further optimizes the perspective of parameter quantity.

5 Conclusion

A remote sensing image road extraction method combining semantic segmentation and angle prediction is proposed in this article, based on the semantic segmentation module, an angle prediction module is added in this method, the main function of which is to predict the tilt angle of the road in the remote sensing image, and use the angle feature fusion module to fuse the two types of features. This can make better use of road angle information and get more accurate road extraction results. The joint semantic segmentation and remote sensing image road extraction algorithm of angle prediction on the Deep Globe dataset is compared with the method based on the fully convolutional neural network, the method based on the Transformer structure, and the road extraction method that has performed well in recent years. The experimental results show that the joint method is superior to the existing semantic segmentation method and road extraction method in quantitative evaluation and visual effect. The research work of this paper also shows that using the road angle as the constraint condition of road extraction tasks can effectively improve the model results and provide reference for related tasks.

The research significance of this paper is to provide a more accurate road region extraction method for high-resolution images, which can obtain better road region extraction results, so as to more accurately obtain the distribution of roads in land use and provide essential data support for studying the quantitative relationship between land subsidence and land use.

The future development trend is mainly focused on the following aspects: the first is to further improve the accuracy and robustness of segmentation by introducing more features and improved algorithms to solve the existing problems; the second is to improve the efficiency and real-time performance of algorithm, so that road semantic segmentation can be more widely used in practical applications; the third is to combine other data sources, such as Lidar and geographic information systems, to further improve the accuracy and comprehensiveness of road semantic segmentation. In general, the research on semantic segmentation and road extraction from remote sensing images is constantly making new breakthroughs, and it is expected to play an important role in traffic planning, intelligent driving, and other fields in the future.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

SX: Conceptualization, Methodology, Validation, Writing–original draft. CM: Conceptualization, Formal Analysis, Resources, Writing–original draft. GY: Validation, Writing–review and editing. YS: Formal Analysis, Writing–review and editing. SL: Methodology, Writing–review and editing. JF: Writing–review and editing.

Funding

The author(s) declare no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Bajcsy, R., and Tavakoli, M. (1976). Computer recognition of roads from satellite pictures. *IEEE Trans. Syst. Man, Cybern.* 9, 623–637. doi:10.1109/tsmc.1976.4309568

Bastani, F., Hesong, T., Abbar, S., et al. (2018). "RoadTracer: automatic extraction of road networks from aerial images," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, June, 2018, 4720–4728.

Carion, N., Massa, F., Synnaeve, G., et al. (2020). "End-to-end object detection with transformers," in Proceedings of the European Conference on Computer Vision, Glasgow, UK, August, 2020, 213–214.

Chen, L. C., Zhu, Y., Papandreou, G., et al. (2018). "Encoder-decoder with aurous separable convolution for semantic image segmen-tation," in Proceedings of the European Conference on Computer Vision, Munich, Germany, September, 2018, 801–818.

Cheng, B., Misra, I., Schwing, A. G., et al. (2021a).Masked-attention mask transformer for universal image segmentation. https://arxiv.org/abs/2112.01527.

Cheng, B., Schwing, A., and Kirillov, A. (2021b). Per-pixel classification is not all you need for semantic segmentation. *Proc. Adv. Neural Inf. Process. Syst.* 74, 17864–17875. doi:10.48550/arXiv.2107.06278

Cheng, G., and Han, J. (2016). A survey on object detection in optical remote sensing images. *ISPRS J. Photogrammetry Remote Sens.* 117, 11–28. doi:10.1016/j.isprsjprs.2016.03.014

Cheng, G., Wang, Y., Xu, S., Wang, H., Xiang, S., and Pan, C. (2017). Automatic road detection and centerline extraction via cascaded end to-end convolutional neural network. *IEEE Trans. Geoscience Remote Sens.* 55 (6), 3322–3337. doi:10.1109/tgrs.2017.2669341

Dai, J. G., Wang, Y., Du, Y., Zhu, T., Xie, S., et al. (2020). Development and prospect of road extraction method for optical remote sensing im-age. *J. Remote Sens.* 24 (7), 804–823. doi:10.11834/jrs.20208360

Deepglobe, .(2023). DEEPGLOBE - CVPR18, http://deepglobe.org/challenge.html.

Demir, I., Koperski, K., Lindenbaum, D., et al. (2018). "DeepGlobe 2018: a challenge to parse the earth through satellite images," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, June, 2018, 172–181.

Deng, Y. L., Yang, M., Zhi, D. L., Yue, S. H., and Wang, C. (2020). Fusing geometrical and visual information via superpoints for the semantic segmentation of 3D road scenes. *Tsinghua Sci. Technol.* 25 (4), 498–507. doi:10.26599/tst.2019.9010038

Drăguț, L., Csillik, O., Eisank, C., and Tiede, D. (2014). Automated parameterisation for multi-scale image segmentation on multiple layers. *ISPRS J. Photogrammetry Remote Sens.* 88, 119–127. doi:10.1016/j.isprsjprs.2013.11.018

Fei, L. W., and Man, C. Y. (2021). Review on semantic segmentation of road scenes. *Laser and Optoelectron. Prog.* 60 (12), 36–58. doi:10.3788/LOP202158.1200002

Herumurti, D., Uchimura, K., Koutaki, G., et al. (2013). "Urban road network extraction based on zebra crossing detection from a very high resolution RGB aerial image and DSM data," in Proceedings of the International Conference on Signal-Image Technology and Internet-Based Systems, Kyoto, Japan, December, 2013, 79–84.

Li, J., Meng, Y., Dorjee, D., Wei, X., Zhang, Z., and Zhang, W. (2021). Automatic road extraction from remote sensing imagery using ensemble learning and post-processing. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* 14, 10535–10547. doi:10.1109/jstars.2021.3094673

Li, R., Liu, W., Yang, L., Sun, S., Hu, W., Zhang, F., et al. (2018). Deepunet: a deep fully convolutional network for pixel-level sea-land segmentation. *Pro-ceedings IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* 11 (11), 3954–3962. doi:10.1109/jstars.2018.2833382

Liu, Z., Lin, Y., Cao, Y., et al. (2021). "Swin transformer: hierarchical vision transformer using shifted windows," in Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, October, 2021, 10012–10022.

Lu, X., Zhong, Y., and Zheng, Z. (2020). A novel global-aware deep network for road detection of very high resolution remote sensing imagery. *Proc. IEEE Int. Geoscience Remote Sens. Symposium*, 2579–2582. doi:10.1109/IGARSS39084.2020.9323155

Lu, X., Zhong, Y., Zheng, Z., and Zhang, L. (2021). GAMSNet: globally aware road detection network with multi-scale residual learning. *ISPRS J. Photogrammetry Remote Sens.* 175, 340–352. doi:10.1016/j.isprsjprs.2021.03.008

Mei, J., Li, R. J., Gao, W., and Cheng, M. M. (2021). CoANet: connectivity attention network for road extraction from satellite imagery. *IEEE Trans. Image Process.* 30, 8540–8552. doi:10.1109/tip.2021.3117076

Rathinam, S., Kim, Z. W., and Sengupta, R. (2008). Vision-based monitoring of locally linear structures using an unmanned aerial vehi-cle. *J. Infrastructure Syst.* 14 (1), 52–63. doi:10.1061/(asce)1076-0342(2008)14:1(52)

RobertCarolaStefan, H. K. H. (2013). Measuring completeness of building footprints in OpenStreetMap over space and time. *ISPRS Int. J. Geo-Information* 2 (4), 1066–1091. doi:10.3390/ijgi2041066

Saito, S., Yamashita, T., and Aoki, Y. (2016). Multiple object extraction from aerial imagery with convolutional neural networks. *Electron. Imaging* 10, 010402-1–010402-9. doi:10.2352/j.imagingsci.technol.2016.60.1.010402

Shao, Z., Tang, P., Wang, Z., Saleem, N., Yam, S., and Sommai, C. (2020). BRRNet: a fully convolutional neural network for automatic building extraction from high-resolution remote sensing images. *Remote Sens.* 12 (6), 1050. doi:10.3390/rs12061050

Sheng, H. L., Huang, P., and Su, Y. (2015). A method for road extraction from remote sensing imagery. *REMOTE SENS-ING LAND and Resour.* 27 (2), 56–62.

Singh, S., Batra, A., Pang, G., et al. (2018). Self-supervised feature learning for semantic segmentation of overhead imagery. *Proc. Br. Mach. Vis. Conf.*, 1–13.

SpaceNetChallenge, .(2023), SpaceNetChallenge, https://github.com/SpaceNetChallenge.

Tan, Y., Gao, S., Li, X., et al. (2020). "VecRoad: point-based iterative graph exploration for road graphs extraction," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, June, 2020, 8907–8915.

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). "Attention is all you need," in Proceedings of the Advances in Neural Infor-mation Processing Systems, Long Beach, California, USA, December, 2017, 5998–6008.

Vosselman, G., and Knecht, J. d. (1995). *Road tracing by profile matching and Kaiman filtering.* Berlin, Germany: Springer, 265–274.

Wan, J., Xie, Z., Xu, Y., Chen, S., and Qiu, Q. (2021). DA-RoadNet: a dual-attention network for road extraction from high resolution satellite im-agery. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* 14, 6302–6315. doi:10.1109/jstars.2021.3083055

Wei, Y., Wang, Z., and Xu, M. (2017). Road structure refined CNN for road extraction in aerial image. *IEEE Geoscience Remote Sens. Lett.* 14 (5), 709–713. doi:10.1109/lgrs.2017.2672734

Willrich, F. (2002). Quality control and updating of road data by GIS-driven road extraction from imagery. *Int. Archives Photogrammetry Remote Sens. Spatial Inf. Sci.* 34 (4), 761–767.

Xin, S., Yi, C., Chuanxin, R., et al. (2022). Influence of pipeline leakage on the ground settlement around the tunnel during shield tunneling. *Sustainability* 14, 14–24.

Yan, M., and Gulimila, K. (2023). Research review of image semantic segmentation method in high-resolution remote sensing image interpretation. *J. Front. Comput. Sci. Technol.* 17 (7), 1526–1548. doi:10.3778/j.issn.1673-9418.2211015

Zainuri, M., Helmi, M., Novita, A. G. M., Pancasakti Kusumaningrum, H., and Koch, M. (2022). An improve performance of geospatial model to access the tidal flood impact on land use by evaluating sea level rise and land subsidence parameters. *J. Ecol. Eng.* 23 (2), 1–11. doi:10.12911/22998993/144785

Zhong, Z., Li, J., Cui, W., et al. (2016). Fully convolutional networks for building and road extraction: preliminary results. *Proc. IEEE Int. Geoscience Remote Sens. Symposium*, 1591–1594. doi:10.1109/IGARSS.2016.7729406

Zhu, Q., Zhang, Y., Wang, L., Zhong, Y., Guan, Q., Lu, X., et al. (2021). A global context-aware and batch-independent network for road extraction from VHR satellite imagery. *ISPRS J. Photogrammetry Remote Sens.* 175, 353–365. doi:10.1016/j.isprsjprs.2021.03.016