Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.048608

ARTICLE

Weakly Supervised Network with Scribble-Supervised and Edge-Mask for Road Extraction from High-Resolution Remote Sensing Images

Supeng Yu[1], **Fen Huang**[1,*] **and Chengcheng Fan**[2,3,4,*]

[1]College of Artificial Intelligence, Nanjing Agricultural University, Nanjing, 210095, China

[2]Innovation Academy for Microsatellites of CAS, Shanghai, 201210, China

[3]Shanghai Engineering Center for Microsatellites, Shanghai, 201210, China

[4]Key Laboratory of Satellite Digital Technology, Shanghai, 201210, China

*Corresponding Authors: Fen Huang. Email: fenhuang@njau.edu.cn; Chengcheng Fan. Email: fancc@microsate.com

Received: 13 December 2023 Accepted: 19 February 2024 Published: 25 April 2024

ABSTRACT

Significant advancements have been achieved in road surface extraction based on high-resolution remote sensing image processing. Most current methods rely on fully supervised learning, which necessitates enormous human effort to label the image. Within this field, other research endeavors utilize weakly supervised methods. These approaches aim to reduce the expenses associated with annotation by leveraging sparsely annotated data, such as scribbles. This paper presents a novel technique called a weakly supervised network using scribble-supervised and edge-mask (WSSE-net). This network is a three-branch network architecture, whereby each branch is equipped with a distinct decoder module dedicated to road extraction tasks. One of the branches is dedicated to generating edge masks using edge detection algorithms and optimizing road edge details. The other two branches supervise the model's training by employing scribble labels and spreading scribble information throughout the image. To address the historical flaw that created pseudo-labels that are not updated with network training, we use mixup to blend prediction results dynamically and continually update new pseudo-labels to steer network training. Our solution demonstrates efficient operation by simultaneously considering both edge-mask aid and dynamic pseudo-label support. The studies are conducted on three separate road datasets, which consist primarily of high-resolution remote-sensing satellite photos and drone images. The experimental findings suggest that our methodology performs better than advanced scribble-supervised approaches and specific traditional fully supervised methods.

KEYWORDS

Semantic segmentation; road extraction; weakly supervised learning; scribble supervision; remote sensing image

1 Introduction

The process of road extraction, which is often referred to as road detection or road segmentation, is a vital undertaking in the fields of computer vision and remote sensing. Autonomous driving, urban planning, and environmental monitoring are among the fields in which it assumes a pivotal role. Road extraction aims to precisely detect and outline road areas in aerial or satellite images. This task can be difficult due to the intricate and varied road configurations in real-life environments. Recently, road extraction tasks have witnessed significant achievements using deep learning-based methods. These methods exploit the capabilities of convolutional neural networks (CNNs) to acquire highly distinctive features from unprocessed visual data [1]. For example, Khan et al. [2] proposed an encoder-decoder network with an integrated attention unit to cope with the road segmentation task in high spatial resolution satellite images, which can automatically analyze high spatial resolution satellite images and extract road networks. Adopting these technologies has significantly improved road extraction precision and efficacy compared to traditional image processing approaches. The expansion of satellite remote sensing technology has dramatically increased the number of open-source roads accessible on OpenStreetMap (OSM). Nonetheless, some locations still need to be explored and undocumented on the global map. Although fully supervised learning may accurately extract road information from remote sensing data, it requires pixel-level labeling, which involves much human effort.

In contrast, weakly supervised learning refers to the process of acquiring knowledge from labels that need to be more sparsely marked. Furthermore, the utilization of weakly supervised learning techniques has the potential to decrease the size of the dataset while simultaneously producing improved classification results.

Standard labeling methods include point annotations [3], scribbles [4], bounding boxes [5], and image-level annotations [6]. Using different types of sparse labels can lead to varying training classification outcomes. Taking scribble annotations as an example, ScribbleSup [4] employs alternate optimization to combine GrabCut [7] and Fully Convolutional Networks (FCN) [8] to enhance segmentation accuracy. However, this approach also increases model complexity and challenges in segmenting fine-edge details. During the network training process, as the similarity between the network's predicted results and the pseudo-labels increases, the supervision obtained from pseudo-labels gradually weakens, and the learning process becomes stable. This phenomenon is referred to as *Lazy Learning* [9]. In *Lazy Learning*, the network model stores knowledge in the pseudo-labels for the student model to learn from. However, if prediction errors occur, they persist throughout the entire self-learning process. As a result, inaccuracies accumulate, ultimately leading to a degradation in the quality of generated pseudo-labels. *Lazy Learning* reflects the lack of quantitative improvement in pseudo-labels during the learning process. The method provided by Sohn et al. [10] involves utilizing consistency regularization to construct pseudo-labels. The initial step consists of generating pseudo-labels for unlabeled images that have undergone weak augmentation. In the context of a specific image, pseudo-labels are selectively preserved solely when the model generates predictions surpassing a pre-determined threshold. After the same picture is significantly augmented, the predictions are changed by computing cross-entropy loss until they match the conserved pseudo-labels. This method successfully reduces pseudo-label imprecision and ensures the integrity of the generated pseudo-labels. Zhang et al. [11] introduced a novel algorithm (Mixup) for image mixup augmentation within the domain of computer vision. The method combines two images from different classes using a random blending strategy to increase the size of the training dataset. Expanding the dataset size may significantly improve the classification accuracy with only a slight increase in Central Processing Unit (CPU) resources.

The application of weakly supervised learning in the domain of remote sensing imagery shows significant potential. Nonetheless, the utilization of weakly supervised deep learning approaches for road surface extraction is still in its early stages due to remote sensing images' numerous and complex characteristics. As a result, exploring methods for efficiently incorporating weak-label information into standard weakly supervised learning algorithms has become a crucial topic that requires careful consideration. Therefore, we propose a novel multi-branch network, known as WSSE-net, that utilizes weak supervision in the form of scribbles and edge masks [12]. This strategy aims to effectively address the challenges related to the complexity of models and the precise segmentation of complicated edge features.

The research conducted has yielded the following primary contributions:

- A novel weakly supervised deep learning strategy has been proposed in this study, which utilizes weak supervision in the form of scribbles and edge masks to extract road information from remote sensing images.
- A convolutional neural network has been developed to produce high-quality pseudo-labels for propagating scribbling annotations. This model integrates edge masks to guide the optimization of road edge information. Mixup dynamically blends prediction results and continually updates new pseudo-labels to steer network training.
- Extensive experimentation has been undertaken using widely recognized public datasets to showcase the efficacy of the strategy described in our study. The system demonstrates exceptional performance and outperforms many widely used scribble-supervised segmentation techniques.

2 Methods

This study presents an innovative approach for road extraction using a weakly supervised convolutional network. Our method incorporates edge masks and scribbling information to enhance the accuracy and efficiency of the road extraction process. Fig. 1 illustrates the model's composition, which comprises a single encoder and three decoders: The primary segmentation decoder, the auxiliary segmentation decoder, and the edge-mask auxiliary decoder. The proposed model utilizes the U-shaped network architecture as the foundational segmentation network and expands it into a three-branch segmentation network by integrating an auxiliary segmentation decoder.

2.1 Edge Mask Auxiliary Decoding Branch

Drawing from the work of Wei et al. [12], it is imperative to integrate high-resolution predictions that exhibit distinct edges with more resilient, lower-resolution features. This amalgamation is crucial for achieving enhanced edge details and minimizing false positives in the segmentation process. Consequently, the information at the lower-resolution level is increased to align with the resolution of the high-resolution information. Subsequently, the Holistically-Nested Edge Detection (HED) [13] technique is employed to merge the abovementioned components, facilitating boundary prediction production. The HED model utilized in this study has undergone pre-training on the Berkeley Segmentation Dataset (BSDS500) [14]. The outlined methodology is as follows: Initially, using HED is employed to generate edge-masks. HED is mainly reformed based on the VGG network. The pooling layer after the fifth convolutional layer of the network is designed by the Visual Geometry Group (VGG). All the fully connected layers are deleted, and the remaining part is used as the primary network. HED employs a stacked structure and a globally integrated perspective to enable simultaneous learning and integration of information at multiple scales. In a road scene, road edge elements typically include objects of varying sizes, such as junctions. HED's multi-scale architecture allows the algorithm to collect more information on the road edge and enhances sensitivity to edges at various scales. HED trains multi-scale edge response graphs simultaneously, allowing the network to interpret the semantic information in pictures better. In the road scene, the road boundary frequently has semantic properties. HED increases the perception of semantic information through collaborative training, allowing for more accurate detection and extraction of features from road edges. The

Edge Mask Auxiliary Decoding Branch also extracts multi-scale features from the Primary Encoder-Decoder Branch.

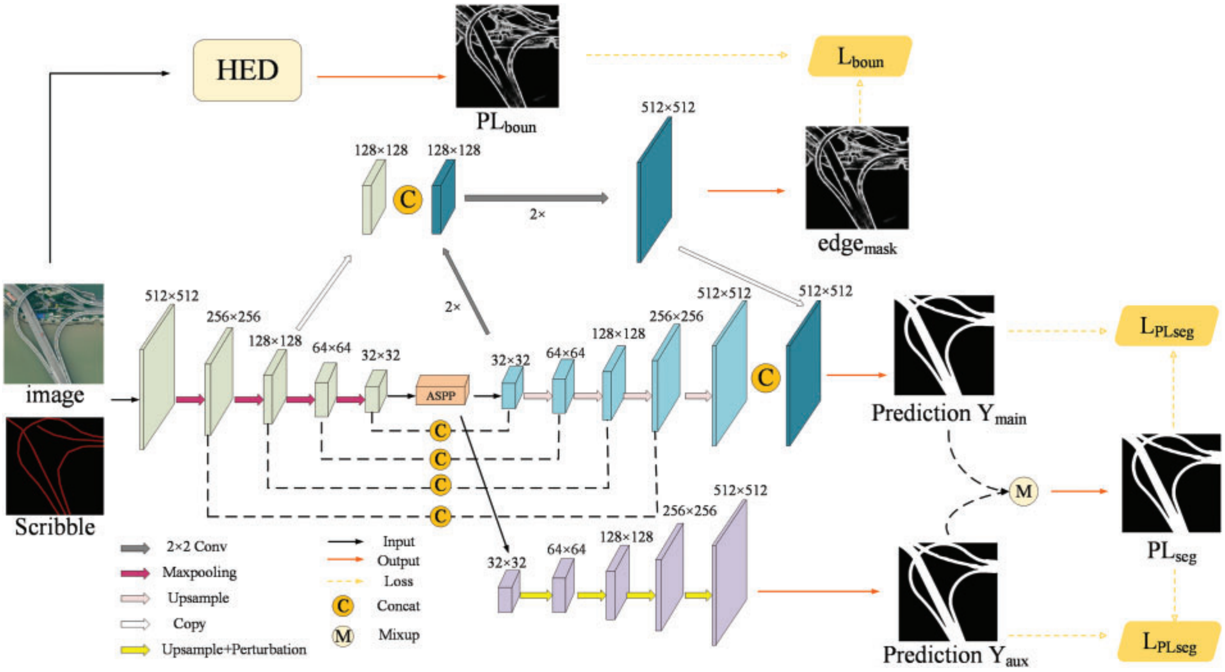

Figure 1: The overall structure of the WSSE-net

The structure of the Edge Mask Auxiliary Decoding Branch: In the Primary Encoder-Decoder Branch, the decoder takes a 32 × 32 pixel feature and performs two separate 2 × 2 up-sampling operations followed by a 3 × 3 convolution operation. The resulting features are then merged with the 128 × 128 pixel encoder. The merged result is then subjected to two additional 2 × 2 up-sampling operations and one 3 × 3 convolution operation to obtain the segmentation prediction $edge_{mask}$. The border pseudo-label is evaluated to analyze the loss in conjunction with the projected outcome $edge_{mask}$. This process aids in the refinement and direction of the significant segmentation branch.

2.2 Auxiliary Decoder Branch

The Auxiliary Decoder Branch predominantly uses the U-net encoder as foundational and introduces a recently defined auxiliary decoder. The decoder comprises four up-sampling layers, and a *dropout* [15] layer is incorporated before each convolutional section in the auxiliary segmentation decoder. Including dropout layers serves the objective of introducing perturbations and enhancing the robustness and generalization power of the model. The segmentation pseudo-label, denoted as PL_{seg}, is created by randomly combining the final generated segmentation prediction, referred to as Y_{aux}, with the primary segmentation prediction, denoted as Y_{seg}.

2.3 Primary Encoder-Decoder Branch

In this paper, we propose a novel method based on Mixup that generates pseudo-labels through random mixing to enhance image segmentation. The method comprises two distinct and independent branches: The Primary Encoder-Decoder Branch and the Auxiliary Decoder Branch. The primary segmentation division employs U-net [16] architecture as the segmentation network's foundational framework. The particular procedure is as follows: Initial input is a 512 × 512 pixel image. The features are then passed through the four down-sampling layers of the decoder's principal segmentation branch. At each down-sampling stage, the feature channels are multiplied by two to reduce the dimensions, improving feature extraction.

The Atrous Spatial Pyramid Pooling (ASPP) module [17] is employed to connect the encoder and decoder. The ASPP module consists of three Atrous convolutions with a size of 3 × 3 and one 1 × 1 convolution. This technique facilitates the acquisition of a broader receptive field while maintaining a substantial level of resolution. In the primary segmentation decoder, the features undergo four up-sampling layers and are merged with $edge_{mask}$ from the auxiliary segmentation decoding branch to produce Y_{seg}. The resultant Y_{seg}, the final output, is combined with Y_{aux}, created from the auxiliary segmentation branch, through random mixup. This process yields pseudo-labels that are utilized for segmentation purposes. The utilization of pseudo-labels serves the purpose of providing additional supervision and training for both the primary segmentation branch and the auxiliary segmentation branch to improve the performance of image segmentation.

Mixup is a mixed-class enhancement algorithm used in computer vision, which can mix images of different classes to expand the training dataset. The core formula of Mixup is as follows:

$$x = \lambda \times x_i + (1 - \lambda) \times x_j \tag{1}$$

$$y = \lambda \times y_i + (1 - \lambda) \times y_j \tag{2}$$

where x_i and x_j are raw input vectors, y_i and y_j are one-hot label encodings. In this paper, we adopted the core idea of mixup as described above and applied it to the model with the following specific formula:

$$PL_{seg} = argmax\left(\lambda \times Y_{seg} + (1 - \lambda) \times Y_{aux}\right) \tag{3}$$

where PL_{seg} is the mixed-generated pseudo-labels, Y_{seg} is the segmentation prediction from the primary segmentation branch, Y_{aux} is the segmentation prediction from the auxiliary segmentation branch, λ is a random number between 0 and 1, which is generated randomly at each iteration.

The argmax function is utilized to get the class ID that corresponds to the highest probability value in the model's prediction. This ensures that each mixing operation yields a more favorable outcome.

In the meantime, scribble trains the segmentation network directly by minimizing the partial cross-entropy loss. This method overcomes the inherent flaw of previous pseudo-label generation, in which the generated labels needed to be updated in conjunction with network training. It accomplishes this by eliminating the gradient between the primary segmentation decoding branch and the auxiliary segmentation decoding branch, thereby preserving their independence instead of enforcing direct consistency. This approach extends the supervisory signal from a limited number of pixels to encompass the entire image. Consequently, the pixels labeled through scribbles can effectively spread across the image by dynamically combining pseudo-labels with the unlabeled pixels [13].

2.4 Optimization Method

As a functional metric that measures the difference between the model outputs and the anticipated labels, the loss function is essential to deep network learning and affects the process's overall efficacy. The loss function makes the network self-optimize by requiring the communication of the computed loss value back to the model. Usually, the application's context and specific criteria influence the choice of loss function.

Based on the two types of labels provided by the network in this paper, namely segmentation pseudo-labels (PL_{seg}) and boundary pseudo-labels (PL_{boun}), the following loss functions are proposed:

$$L_{boun} = \sum_{i=1}^{w} \sum_{j=1}^{h} \left(PL_{boun} - edge_{mask}\right)^2 \tag{4}$$

The equation includes the boundary loss, where w and h represent the width and height of the boundary pseudo-labels (PL_{boun}) at the pixel level, and $edge_{mask}$ represents the prediction of the edge mask auxiliary decoding branch. The segmentation pseudo-labels generated (PL_{seg}) are utilized to compute the loss function independently for the segmentation predictions of both the primary decoder branch (Y_{seg}) and the auxiliary decoder branch (Y_{aux}). The specific loss function can be expressed as follows:

$$L_{PL_{seg}} = CE\left(Y_{seg}, PL_{seg}\right) + CE\left(Y_{aux}, PL_{seg}\right) \tag{5}$$

The segmentation prediction Y_{seg} is only used to calculate the loss function for the scribble-labeled regions, rather than the entire image. In the scribble regions, the local loss function is defined as follows:

$$L_{scri} = \sum_{i \in W_m} \log Y_{segi} + \sum_{i \in W_m} \log Y_{auxi} \tag{6}$$

In the equation, m represents the one-hot encoded scribble-labeled regions, which indicate the probability of pixel i belonging to the road. W_m denotes the set of scribble-labeled pixels. Based on the above information, the overall loss function can be formulated as follows:

$$L_{final} = \alpha L_{boun} + \gamma L_{scri} + L_{PL_{seg}} \tag{7}$$

2.5 Dataset Descriptions

To evaluate the performance of the proposed method, we validated our method on three road datasets, including the CHN6-CUG dataset [18,19], the Berlin Road dataset [20], and the Massachusetts Dataset. The CHN6-CUG road dataset primarily covers urban areas in China, including six Chinese cities, such as Chaoyang District in Beijing, Yangpu District in Shanghai, and the city center of Wuhan. It is a pixel-level dataset containing road and non-road data. CHN6-CUG contains 4511 annotated images with a size of 512 × 512 pixels. These images are divided into 3608 for model training and 903 for testing and evaluation. Kaiser created the Berlin Road dataset from high-resolution satellite imagery. The collection primarily focuses on the urban areas of Berlin. It is composed of aerial images from Google Earth, as well as pixel-level building, road, and background labels from OpenStreetMap. We performed post-processing on the images and labels of this dataset, preserving road centerlines as scribble annotations, resulting in the Berlin Road Dataset. The Massachusetts Road dataset consists of 1171 aerial images from Massachusetts. As with building data, the size of each image is 1500 × 1500 pixels, covering an area of 2.25 square kilometers. We randomly divided the data into a training set of 1080 images, a validation set of 21, and a test set of 70 images. These three datasets have undergone post-processing scribble annotation. These three data sets are all correlated. They are all road data sets based on remote sensing images taken by satellites, and they all cover various roads in different scenes, such as cities, rural areas, and mountain areas. However, there are some differences: The Berlin dataset focuses more on road occlusion in the city, etc., and the CHN6-CUG includes many mountains, rural, and field roads, focusing on the edge details of the road, while the Massachusetts dataset focuses on road continuity. The first dataset uses line annotations from LabelMe as scribbles (Fig. 2), while the second dataset uses road centerlines from OpenStreetMap (OSM) as scribbles.

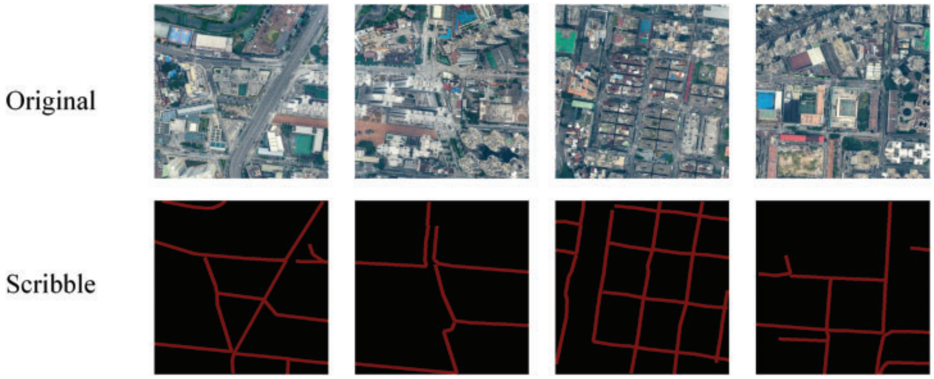

Figure 2: Examples of scribble annotations

2.6 Experimental Details

The proposed WSSE-net was implemented using Pytorch11.6 on NVIDIA RTX 2070 GPU. All our experiments are based on the same hardware and software platform, different hardware and software devices may have other effects on the experimental results. The network performance parameters of WSSE-net are evaluated on the test set using multiple evaluation metrics. We standardized the datasets to a size of 512 × 512 for the experiments and then restored them to their original sizes at the end of the experiment.

2.7 Evaluation Metrics

Precision, recall, F1 Score (also known as F1-Score or F1), and intersection over union (IoU) are widely used for assessing pixel-level segmentation. Precision is a metric that quantifies the ratio of accurately predicted pixels to the total number of anticipated pixels. These four measures were chosen to assess the performance of the proposed WSSE-net.

First, we calculate the confusion matrix for road predictions and road ground truth, including True-Positive (TP), False-Positive (FP), False-Negative (FN) and True-Negative (TN). Precision, recall, F1 Score, and IoU can be calculated as follows:

$$IoU = \frac{TP}{TP + FN + FP} \tag{8}$$

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

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

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{11}$$

3 Results

A series of experiments were undertaken using identical datasets and network settings to assess the validity of the model provided in this research study. A comparative analysis was conducted between the suggested model and a well-recognized Scribble weakly supervised segmentation model as well as WSSE-net. Specifically, the comparison was performed with ScribbleSup [4], ScRoadExtractor [12], BoxSup [21], Boundary Perception Guidance (BPG) [22], Weakly-Supervised Salient Object Detection via Scribble Annotations (WSOD) [23], and Weakly labeled OpenStreetMap Centerline (WeaklyOSM) [24] models. Additionally, we assess the efficacy of our model by doing a comparative analysis with established fully supervised networks, namely U-net, DeepLabV3+ [17], SegNet [25] and D-LinkNet [26] series. A visual analysis of the segmentation results and quantitative assessment metrics were used to evaluate each network's segmentation performance.

3.1 Results on the CHN6-CUG Dataset

The experimental findings obtained from training samples on the CHN6-CUG dataset are shown in Table 1. The first column of Table 1 presents other weakly supervised techniques compared to WSSE-net. The following four columns provide the outcomes for four distinct assessment measures, namely Precision, Recall, F1 Score (sometimes referred to as F1-Score or F1), and Intersection over Union (abbreviated as IoU). The performance of several approaches on the test dataset is shown in Fig. 3, which exhibits a random selection of three samples from the test dataset for visualization purposes.

Table 1: Quantitative results for WSSE-net and the comparison methods on the CHN6-CUG dataset

Method	IOU/%	F1-Score/%	Precision/%	Recall
ScribbleSup	43.16	59.84	61.09	58.64
BPG	48.15	64.60	74.14	57.23
WSOD	45.75	59.39	68.82	52.24
WeaklyOSM	50.21	67.29	72.45	62.82
ScRoadExtractor	53.40	70.51	**74.96**	66.56
Ours	**55.26**	**70.94**	74.58	**67.54**

3.2 Results on the Berlin Road Dataset

Similarly, Table 2 presents the experimental results using training samples on the Berlin dataset. Fig. 4 demonstrates the efficacy of various methods by randomly selecting three samples from the test dataset using the same process.

3.3 Results on the Massachusetts Dataset

In this investigation, a distinct set of comparative methodologies were utilized than in the previous two. We utilized established, fully supervised procedures. Table 3 presents the results of experiments conducted using training samples on the Massachusetts Road dataset. In particular, we replaced the fully supervised comparative methods in the first column while maintaining the evaluation metrics from the previous experiments. Fig. 5 depicts the efficacy demonstrated by various experimental methodologies.

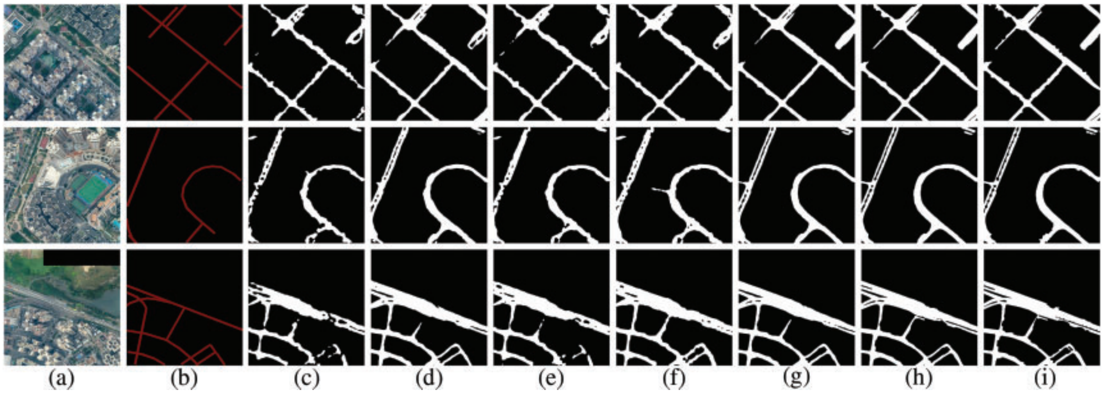

Figure 3: Qualitative results of road segmentation using different methods on the CHN6-CUG dataset. (a) Image. (b) Scribble annotation. (c) ScribbleSup. (d) BPG. (e) WSOD. (f) WeaklyOSM. (g) ScRoadExtractor. (h) Ours. (i) Ground truth

Table 2: Quantitative results for WSSE-net and the comparison methods on the Berlin Road dataset

Method	IOU/%	F1-Score/%	Precision/%	Recall
ScribbleSup	50.76	60.39	65.66	55.92
BPG	59.84	65.66	72.24	60.18
WSOD	56.55	62.71	70.05	56.77
WeaklyOSM	62.42	67.48	**75.18**	61.21
ScRoadExtractor	64.25	69.42	73.86	65.49
Ours	**65.02**	**70.10**	74.52	**66.19**

Table 3: Quantitative results for WSSE-net and the other full-supervised methods on the Massachusetts Road dataset

Method	IOU/%	F1-Score/%	Precision/%	Recall
U-Net	55.10	70.57	76.54	65.46
SegNet	56.98	71.74	77.93	66.42
DeepLabV3+	62.37	76.86	75.86	77.89
D-LinkNet34	61.35	75.78	72.64	79.20
D-LinkNet50	**63.07**	**77.32**	73.45	**81.64**
Ours	61.88	77.22	**79.67**	74.92

3.4 Ablation Studies

This section conducts a series of ablation experiments using the CHN6-CUG dataset to assess the performance of the different modules inside the WSSE-net architecture. In the ablation experiments of the loss function, we performed three comparison experiments. The first comparative experiment focused only on selecting the segmentation loss (L_{PLseg}). In the subsequent comparative experiment, we included both the segmentation loss (L_{PLseg}) and the Scribble label loss (L_{scri}). The third comparative experiment included the edge loss (L_{boun}), building upon the methodology of the second experiment. We conducted additional ablation trials in the edge identification section while preserving the integrity of the overall loss function. In these trials, the edge detection techniques used were *HED*, *Canny* [27], and *DeepEdge* [28], each applied individually. The detailed findings may be seen in Tables 4 and 5.

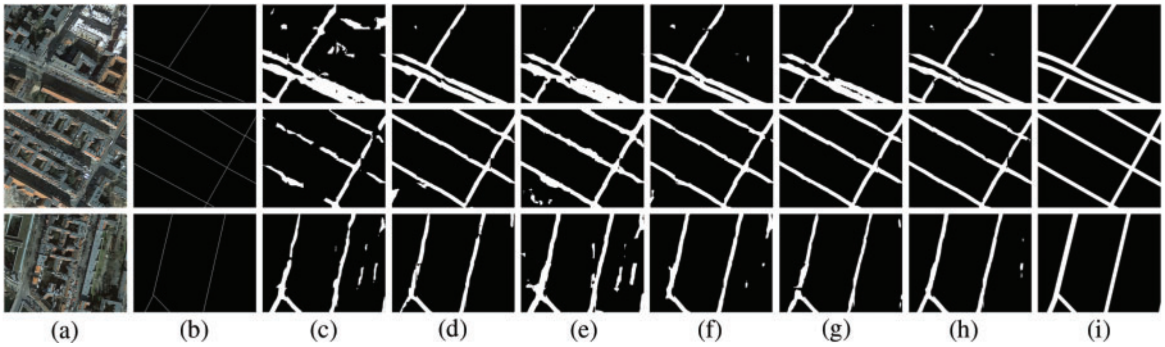

Figure 4: Qualitative results of road segmentation using different methods on the Berlin Road dataset. (a) Image. (b) Scribble annotation. (c) ScribbleSup. (d) BPG. (e) WSOD. (f) WeaklyOSM. (g) ScRoadExtractor. (h) Ours. (i) Ground truth

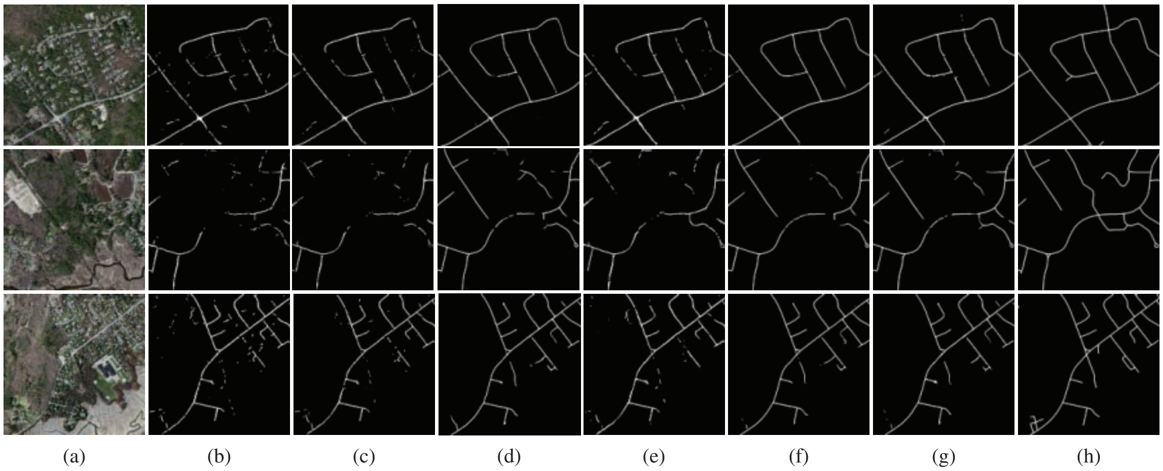

Figure 5: Qualitative results of road segmentation using different methods on the Massachusetts Dataset. (a) Image. (b) U-Net. (c) SegNet. (d) DeepLabV3+. (e) D-LinkNet34. (f) D-LinkNet50. (g) Ours. (h) Ground truth

Table 4: Comparison of segmentation accuracy with different loss functions

Method	IOU/%	F1-Score/%	Precision/%	Recall
L_{PLseg}	49.27	62.55	67.44	60.09
$L_{PLseg} + L_{scri}$	51.31	66.04	70.35	62.24
L_{final}	**54.76**	**68.71**	**72.15**	**65.58**

Table 5: Comparison of segmentation accuracy with different edge mask methods

Method	IOU/%	F1-Score/%	Precision/%	Recall
Canny	49.78	65.07	70.84	60.18
DeepEdge	52.10	68.44	**73.37**	64.84
HED	**54.83**	**69.25**	72.70	**66.12**

4 Discussion

The performance of our method surpasses that of current advanced weakly supervised methods. Upon careful examination of the experimental findings, a comprehensive evaluation of several weakly supervised models is shown in Table 1 and Fig. 3. The model that exhibits the highest level of performance is emphasized via the use of bold formatting. The results indicate that ScribbleSup showed subpar performance, displaying notably worse results than BPG, WSOD, WeaklyOSM, ScRoadExtractor, and our WSSE-net approaches. One limitation of BPG is its limited interaction between the split branch and the border branch, which occurs only at the level of the loss function. This approach does not consider the inherent connection between the two subbranches. Based on the findings shown in Table 1, it can be inferred that the WSSE-net model we have suggested has superior performance in terms of recall (0.6754), IoU (0.5526), and F1 Score (0.7094) when evaluated on the CHN6-CUG dataset. Regarding recall, WSSE-net exhibits a 4.72% enhancement compared to WeaklyOSM, suggesting that our model can extract road network topologies that are more comprehensive. Compared to ScRoadExtractor, our Intersection over Union (IoU) metric has seen a significant rise of 1.96%. This finding implies that the road labels retrieved by WSSE-net exhibit higher alignment with the ground truth data. In the second experiment, as shown in Table 2, WeaklyOSM exhibits an accuracy marginally greater than WSSE-net by 0.66%. However, our model demonstrates a notable enhancement of around 3% in terms of IOU and F1 Score, suggesting a superior quality in comparison. There is no significant difference in the assessment metrics between ScRoadExtractor and our method. However, while analyzing Fig. 4, it becomes evident that the segmentation performance of ScRoadExtractor is comparatively worse than that of our technique. Our technique outperforms ScRoadExtractor in terms of road connectivity and completeness. Our technique also performs well when compared to existing popular full supervision methods. In the third experiment, a series of comparison tests were done on the Massachusetts road dataset. The objective was to evaluate the performance of our technique with various well-established, fully supervised semantic segmentation methods, including U-Net, SegNet, DeepLabv3+, and the D-LinkNet family of approaches. The backbone of D-LinkNet34 is built upon a ResNet34 pre-trained on ImageNet, while the backbones of DeepLabv3+ and D-LinkNet50 are pre-trained on ImageNet using ResNet50. Based on the data shown in Table 3, it is apparent that our technique demonstrates a significant superiority over U-Net and SegNet in relation to IOU while somewhat falling behind DeeplabV3+ and D-LinkNet50. Regarding the F1 Score metric, our approach demonstrates a marginal difference of just 0.1% compared to D-linkNet. Furthermore, our technique exhibits superior performance in terms of accuracy when compared to other methodologies. As seen in Fig. 5, the outcomes of our segmentation analysis indicate that our weakly supervised segmentation method based on scribbles has comparable efficacy to some conventional fully supervised models.

From the ablation experiments on the loss functions, while maintaining the other network configurations constant, the loss functions are configured as follows: Using only the segmentation loss(L_{PLseg}), utilizing both the segmentation loss(L_{PLseg}) and Scribble label loss(L_{scri}), and employing a composite loss function that integrates segmentation loss(L_{PLseg}), Scribble label loss(L_{scri}), and edge loss(L_{boun}). The experimental outcomes, as illustrated in Table 4, demonstrate that the employed loss function of the model achieves an Intersection over Union (IoU) of 54.76%, an accuracy of 72.15%, an F1 Score of 68.71%, and a recall rate of 65.58%. The assessment results for the suggested integrated loss function demonstrate its superiority over a given individual loss function. This finding suggests that the composite loss function outperforms the loss function when applied to a complicated dataset. The ablation tests in the edge mask section show that the edge branch's inclusion may effectively improve the precision of edge details. This improvement is achieved by integrating high-resolution predictions of unique edges with more resilient, lower-resolution features, correcting false positive detections. In this study, we compared the edge pseudo-label generation component, denoted as *HED*, and two standard edge creation algorithms, namely *Canny* and *DeepEdge*. The outcomes of this comparison are shown in Table 5. The detection performance of the *HED* algorithm exhibits a modest boost compared to the *DeepEdge* and *Canny* algorithms.

5 Conclusion

This paper introduced a weakly supervised image segmentation method based on scribble-assisted supervision and edge-masks assistance. The methodology used in this study involves implementing a multi-branch convolutional neural network to conduct end-to-end training. Expanding upon the multi-branch network, we proposed strategies for edge-mask guided assistance and random dynamic mixing to generate pseudo-labels for aiding the training process. Experiments were done on the CHN6-CUG and Berlin Road datasets to validate the efficacy of the suggested methodology. Furthermore, our technique outperformed five traditional fully supervised segmentation algorithms when evaluated on the public Massachusetts road dataset. In this study, high-resolution remote sensing image road surface extraction is realized, which effectively reduces the cost of manual labeling and is ahead of other weakly supervised learning methods. Compared with fully supervised learning, this algorithm can improve the automation of road extraction, but its performance still lags that of the popular fully supervised learning. Weak supervision algorithms may face challenges when processing complex objects or images containing fine-grained structures. Therefore, combining deep learning with traditional methods as a breakthrough to mine potential supervision signals of unlabeled image data without requiring a lot of manual annotation work will be an essential step in the research of intelligent road extraction of remote sensing images.

In the future, we are going to keep working on our study on road extraction inside the region of inadequate supervision, and we are going to keep developing our methodology. In addition, one of our goals is to apply and evaluate our suggested technique to additional complex remote sensing image segmentation tasks.

Acknowledgement: The authors want to thank the School of Artificial Intelligence at Nanjing Agricultural University (Nanjing, China) for providing computing resources to develop this study.

Funding Statement: This work was partly supported by the National Natural Science Foundation of China (42001408, 61806097).

Author Contributions: The authors confirm contribution to the paper as follows: Study conception and design: S. Yu and F. Huang; data collection: S. Yu; analysis and interpretation of results: S. Yu and

C. Fan; draft manuscript preparation: S. Yu and C. Fan. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The related datasets in this article are open source. The dataset with scribbled labels processed in the experiment can be obtained by contacting the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

[1] Q. Zhu, Y. Yang, X. Sun, and M. Guo, "CDANet: Contextual detail-aware network for high-spatial-resolution remote-sensing imagery shadow detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, no. 4, pp. 1–15, 2022. doi: 10.1109/TGRS.2022.3230829.

[2] S. Khan, L. Alarabi, and S. Basalamah, "DSMSA-Net: Deep spatial and multi-scale attention network for road extraction in high spatial resolution satellite images," *Arab. J. Sci. Eng.*, vol. 48, no. 2, pp. 1907–1920, 2023. doi: 10.1007/s13369-022-07082-z.

[3] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei, "What's the point: Semantic segmentation with point supervision," arXiv preprint arXiv:1506.02106, 2015.

[4] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "ScribbleSup: Scribble-supervised convolutional networks for semantic segmentation," in *2016 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 3159–3167. doi: 10.1109/CVPR.2016.344.

[5] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele, "Simple does it: Weakly supervised instance and semantic segmentation," arXiv preprint arXiv:1603.07485, 2016.

[6] J. Ahn and S. Kwak, "Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation," arXiv preprint arXiv:1803.10464, 2018.

[7] C. Rother, V. Kolmogorov, and A. Blake, "GrabCut: Interactive foreground extraction using iterated graph cuts," in *ACM SIGGRAPH 2004 Papers*, New York, NY, USA, 2004, pp. 309–314. doi: 10.1145/1186562.1015720.

[8] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 7–12, 2015, pp. 3431–3440.

[9] X. Huo *et al.*, "ATSO: Asynchronous teacher-student optimization for semi-supervised image segmentation," in *2021 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 20–25, 2021, pp. 1235–1244.

[10] K. Sohn *et al.*, "FixMatch: Simplifying semi-supervised learning with consistency and confidence," arXiv preprint arXiv:2001.07685, 2001.

[11] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," arXiv preprint arXiv:1710.09412, 2017.

[12] Y. Wei and S. Ji, "Scribble-based weakly supervised deep learning for road surface extraction from remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022. doi: 10.1109/TGRS.2021.3061213.

[13] S. Xie and Z. Tu, "Holistically-nested edge detection," in *2015 IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 7–13, 2015, pp. 1395–1403.

[14] P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, 2011. doi: 10.1109/TPAMI.2010.161.

[15] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," arXiv preprint arXiv:1505.04597, 2015.

[17] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," arXiv preprint arXiv:1802.02611, 2018.

[18] Q. Zhu, Z. Li, Y. Zhang, and Q. Guan, “Building extraction from high spatial resolution remote sensing images via multiscale-aware and segmentation-prior conditional random fields,” *Remote Sens.*, vol. 12, pp. 23, 2020. doi: 10.3390/rs12233983.

[19] Q. Zhu *et al.*, “A global context-aware and batch-independent network for road extraction from VHR satellite imagery,” *ISPRS J. Photogramm. Remote Sens.*, vol. 175, no. 12, pp. 353–365, 2021. doi: 10.1016/j.isprsjprs.2021.03.016.

[20] P. Kaiser, J. D. Wegner, A. Lucchi, M. Jaggi, T. Hofmann and K. Schindler, “Learning aerial image segmentation from online maps,” *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 11, pp. 6054–6068, 2017. doi: 10.1109/TGRS.2017.2719738.

[21] J. Dai, K. He, and J. Sun, “BoxSup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation,” arXiv preprint arXiv:1503.01640, 2015.

[22] B. Wang *et al.*, “Boundary perception guidance: A scribble-supervised semantic segmentation approach,” in *Proc. Twenty-Eighth Int. Joint Conf. Artif. Intell.*, Macao, China, Aug. 10–16, 2019. doi: 10.24963/ijcai.2019/508.

[23] J. Zhang, X. Yu, A. Li, P. Song, B. Liu, and Y. Dai, “Weakly-supervised salient object detection via scribble annotations,” arXiv preprint arXiv:2003.07685, 2020.

[24] S. Wu, C. Du, H. Chen, Y. Xu, N. Guo and N. Jing, “Road extraction from very high resolution images using weakly labeled openstreetmap centerline,” *ISPRS Int. J. Geo-Inf.*, vol. 8, no. 11, pp. 478, 2019. doi: 10.3390/ijgi8110478.

[25] V. Badrinarayanan, A. Kendall, and R. Cipolla, “SegNet: A deep convolutional encoder-decoder architecture for image segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, 2017. doi: 10.1109/TPAMI.2016.2644615.

[26] L. Zhou, C. Zhang, and M. Wu, “D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction,” in *2018 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Salt Lake City, UT, USA, Jun. 18–22, 2018, pp. 192–1924.

[27] J. Canny, “A computational approach to edge detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, no. 6, pp. 679–698, 1986. doi: 10.1109/TPAMI.1986.4767851.

[28] G. Bertasius, J. Shi, and L. Torresani, “DeepEdge: A multi-scale bifurcated deep network for top-down contour detection,” arXiv preprint arXiv:1412.1123, 2014.