This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Scribble-Based Weakly Supervised Deep Learning for Road Surface Extraction From Remote Sensing Images

Yao Wei and Shunping Ji, *Member, IEEE*

***Abstract*— Road surface extraction from remote sensing images using deep learning methods has achieved good performance, while most of the existing methods are based on fully supervised learning, which requires a large amount of training data with laborious per-pixel annotation. In this article, we propose a scribble-based weakly supervised road surface extraction method named ScRoadExtractor, which learns from easily accessible scribbles such as centerlines instead of densely annotated road surface ground truths. To propagate semantic information from sparse scribbles to unlabeled pixels, we introduce a road label propagation algorithm, which considers both the buffer-based properties of road networks and the color and spatial information of super-pixels, to produce a proposal mask with categories road, nonroad, and unknown. The proposal mask, along with the auxiliary boundary prior information detected from images, is utilized to train a dual-branch encoder–decoder network which we designed for precise road surface segmentation. We perform experiments on three diverse road data sets that are comprised of high-resolution remote sensing satellite and aerial images across the world. The results demonstrate that ScRoadExtractor exceeds the classic scribble-supervised segmentation method by 20% for the intersection over union (IoU) indicator and outperforms the state-of-the-art scribble-based weakly supervised methods at least 4%.**

***Index Terms*— Remote sensing image, road surface extraction, semantic segmentation, scribble, weakly supervised learning.**

I. Introduction

As A fundamental and important problem in the field of remote sensing image processing, road extraction has a great number of applications including navigation, geo-information database updating, disaster management, and autonomous driving and also provides contextual information that benefits other related tasks, such as land cover classification and vehicle detection. In general, road extraction from remote sensing imagery falls into two subtasks: road surface extraction [1]–[5] and road centerline extraction [6]–[10]. The former focuses on extracting the complete road surfaces from backgrounds as a binary segmentation map, whereas the latter aims to extract the topological structure of road networks in vector form without road width information. With the rapid development of volunteered geographic information (VGI) sources which allow millions of contributors to create and edit geographic data across the world, more road vector data are becoming publicly available. However, these open-source maps are often noisy and/or incomplete. For example, OpenStreetMap (OSM) [11], which is one of the most extensive VGI sources, provides only the coordinates of road centerlines without road width information. Therefore, new effective methods are expected to achieve accurate and automatic road surface extraction from remote sensing images and the existing OSM centerlines with minimum human cost.

Recently, the deep learning [especially deep convolutional neural network (DCNN)]-based road surface extraction methods have been widely studied and achieved good performance. Mnih and Hinton [1] adopted a deep belief network composed of restricted Boltzmann machines (RBMs) for road surface extraction. Panboonyuen *et al.* [12] employed SegNet [13], a variant of fully convolutional network (FCN), in road segmentation and implemented postprocessing using landscape metrics and conditional random field (CRF) [14]. Zhang *et al.* [15] proposed an improved DCNN, which combined ResNet [16] and U-Net [17] to extract roads from remote sensing images. Xu *et al.* [18] utilized a global and local attention model based on U-Net and DenseNet [19]. Zhou *et al.* [20] proposed D-LinkNet model for the road segmentation task, which combined the benefits of encoder–decoder architecture and dilated convolution to capture multiscale features. He *et al.* [21] improved the performance of road extraction models by integrating the Atrous spatial pyramid pool (ASPP) [22] with an encoder–decoder network to enhance the ability of extracting the detailed features of the road. To strengthen the spatial consistency of road segmentation, Zhang *et al.* [23] developed an ensemble strategy by leveraging different FCNs with a weighted loss function. Wei *et al.* [24] proposed a DCNN-based framework that aggregated the semantic and topological information of roads to produce refined road segmentation maps with better connectivity and completeness.

However, training such networks relies on large amounts of densely annotated labels for optimizing millions of parameters. The data-driven-supervised learning approaches may be not practical for industrial applications due to the lack of perfect

Manuscript received October 24, 2020; revised December 23, 2020 and January 29, 2021; accepted February 18, 2021. This work was supported by the National Key Research and Development Program of China, Grant No. 2018YFB0505003. *(Corresponding author: Shunping Ji.)*

The authors are with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: weiyao@whu.edu.cn; jishunping@whu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TGRS.2021.3061213.

Digital Object Identifier 10.1109/TGRS.2021.3061213

0196-2892 © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Wuhan University. Downloaded on March 09,2021 at 14:59:38 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

supervision. To avoid the demand of labor-intensive and time-consuming work, weakly supervised learning, which attempts to learn from low-cost sparse annotation (e.g., scribble, click), has drawn increasingly more attention in computer vision fields. This motivates the road surface extraction to be formulated as a weakly supervised deep learning task.

In this article, we investigate the possibility of training a weakly supervised deep learning model with existing and easily accessible scribbles such as road centerlines derived from manual delineation, OSM data, or GPS traces. We attempt to extract road surface through learning from the scribble annotations, making it possible to greatly reduce the annotation effort. Apparently, this consideration is highly related to the development of VGI sources as noted above and the lack of perfect supervision in real-world applications.

The main contributions of our research are as follows.

1) We propose a novel scribble-based weakly supervised deep learning approach (called ScRoadExtractor) for road surface extraction from remote sensing images under the weak supervision of centerline-like scribble annotations.
2) A road label propagation algorithm is proposed to propagate the semantic information from scribbles to unlabeled pixels by utilizing both the buffer-based properties of roads and the local and global dependencies between graph nodes built on super-pixels to generate the proposal mask.
3) We design a dual-branch encoder–decoder network (DBNet), which is trained with the proposal mask and boundary prior information detected from images, and outputs a road surface segmentation map that approaches to a map predicted from a densely supervised method.
4) The experimental results on diverse road data sets across the world demonstrate that our approach possesses high-performance and powerful generalization ability and also outperforms state-of-the-art scribble-supervised segmentation methods.

The remainder of this article is arranged as follows. In Section II, we briefly review the related studies. Section III provides a detailed description of ScRoadExtractor, including the road label propagation algorithm and the DBNet. Section IV presents the experiments we conducted to verify the effectiveness and generalization ability of the proposed method on diverse data sets in comparison with most recent studies. Discussions are given in Section V, and Section VI presents our conclusions and future research prospects.

II. Related Work

In this section, we briefly review the generic scribble-based weakly supervised methods, and a few weakly supervised learning methods designed for road surface extraction.

As most of the existing deep learning-based road surface extraction methods [12], [15], [18], [20], [21], [23], [24] are fully supervised, per-pixel annotations of road surfaces have to be prepared as training samples, which demands laborious work and is hardly met in practice. In order to tackle this issue, weakly supervised learning has been explored to avoid annotating huge amounts of training data, which involves learning from weak supervision such as scribble [25], [26], click [27], bounding box [28], [29], and image-level tags [30], [31].

Typically, weakly supervised learning methods adopt an alternative training scheme: generate pseudo-semantic labels (i.e., proposals) from the seeds provided by sparse annotated data; train learning models (e.g., DCNNs) with these proposals using standard loss functions (e.g., cross-entropy); and alternate between the proposal generation and the model training steps. Based on scribble annotations and graph theory [32], [33], Lin *et al.* [25] alternated between generating proposals using a graph defined over super-pixels and training an FCN with the proposals. However, they assumed that the labels of super-pixels were constant, which unavoidably led to an artificial upper bound on the accuracy of proposal generation. Similarly, Papandreou *et al.* [34] alternated between two steps in an iterative manner: estimating the latent pixel labels through improved expectation-maximization methods from bounding box annotations and image-level annotations, and training DCNNs in weakly supervised and semi-supervised settings.

Although the quality of the generated proposals can be enhanced by alternating optimization, the two approaches share a common drawback that the training tends to be vulnerable to inaccurate intermediate proposals which are treated as labels, since standard loss functions do not distinguish the seeds from the mislabeled pixels.

Hence, additional regularizations including graph-based approaches (e.g., CRF) and boundary-based losses [35] are often employed to enrich the semantic information of weak annotations in an end-to-end manner. Tang *et al.* [36], [37] introduced a normalized cut loss and a partial cross-entropy loss for weakly supervised semantic segmentation, and incorporated standard regularization techniques (graph cuts and dense CRFs) into the loss function over the partial inputs. Obukhov *et al.* [38] proposed a gated CRF loss for unlabeled pixels together with partial cross-entropy loss for labeled pixels. Kolesnikov and Lampert [35] applied a constrain-to-boundary principle to recover detail information for weakly supervised segmentation. More recently, boundaries have been directly embedded into segmentation network. Wang *et al.* [39] designed a network architecture that consisted of two subnetworks: the prediction refinement network (PRN) and the boundary regression network (BRN), where the BRN guided the PRN in localizing the boundaries. However, the semantics and boundaries information interact only at the loss functions, without considering the relationships between features of the two subnetworks. Zhang *et al.* [40] proposed a weakly supervised salient object detection method by introducing an auxiliary edge detection network and a gated structure-aware loss which focused on the salient regions of images. However, this method was not very stable due to its over-confident saliency predictions.

The aforementioned weakly supervised learning methods had yet to be recognized as promising for road surface extraction until recently. Several studies [41], [42] attempted to employ publicly available maps (e.g., OSM vector data) for road surface extraction from aerial images; however, these

Authorized licensed use limited to: Wuhan University. Downloaded on March 09,2021 at 14:59:38 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

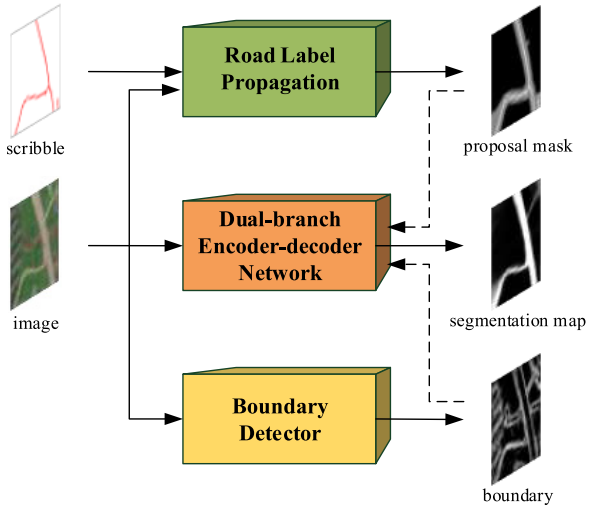


Fig. 1. Overview of the proposed ScRoadExtractor in training phase. The dashed line represents the constraints in the loss function.

methods did not utilize the recent weakly supervised deep learning techniques as mentioned above, and only considered the VGI-based road segmentation as a width estimation problem, making it fragile to complex scenes in high-resolution remote sensing images.

In terms of road vector data, pixels on the road vectors can provide weak supervision for segmenting road surface, which motivates road extraction being formulated as a weakly supervised learning task by two most recent studies. Kaiser *et al.* [43] trained a CNN for building and road extraction with noisy labels generated from OSM data; specifically, the road labels were simply determined by an average road width for each category (e.g., highway, motorway) which was provided by OSM. According to a predefined road width, Wu *et al.* [44] utilized OSM centerline to produce initial road annotation masks, which were then fed into a road segmentation network with a partial loss for labeled pixels and a normalized cut loss for unlabeled pixels. It is obvious that the main shortcoming of the two methods is the fixed road width supervision. This simple and empirical strategy only deals with specific data and has limitations on large-scale and complex data sets. Thus, there are still many more aspects that should be investigated in more detail for weakly supervised deep learning-based road surface extraction.

III. Methodology

We propose ScRoadExtractor, a novel scribble-based weakly supervised deep learning method for road surface extraction from remote sensing images, and the framework is illustrated in Fig. 1. First, the proposal masks are generated by a road label propagation algorithm based on the remote sensing images and scribble annotations. Then, the generated proposal masks and boundary prior information detected from images are used to train a DBNet for road surface extraction by minimizing a joint loss function.

A. Road Label Propagation

Since scribble annotations provide sparse information that limits the overall accuracy of labeling, directly training a DCNN model with sparse scribbles inevitably leads to poor identification results. When taking road centerlines derived from GPS traces or OSM maps as scribbles, a straightforward method is to expand the centerline with a predefined width, but it cannot perfectly identify road boundaries as road width varies. Another possible solution is to mark pixels with features similar to road pixels as roads. Starting from the two straight ideas, we develop a more sophisticated context-aware road label propagation algorithm which propagates the semantic labels from scribbles to unlabeled pixels and marks every pixel of an image within two categories: known (road and nonroad) and unknown pixels, as shown in Fig. 2.

First, considering the property of road networks that the boundaries tend to be parallel to the road centerline, a buffer-based strategy is applied to infer buffer-based masks according to the distance from the road centerline. Specifically, two buffers of scribbles are created with buffer width a_1 and a_2 ($a_1 < a_2$), respectively. The pixels within the first buffer are denoted as the road pixels, the pixels outside the second buffer represent the nonroad pixels, and the remaining pixels are categorized as unknown pixels. However, buffer-based strategy only generates coarse masks and relies on the quality of the scribbles; for example, incorrect and incomplete centerlines may exist in outdated GIS maps or OSM data.

Second, a graphical model is constructed on the super-pixels of a training image by minimizing an energy function, which was inspired by Graph Cut [32], [33] that leverages the unary and pairwise potentials to model the local and global dependencies between graph nodes. First, we employ the simple linear iterative clustering (SLIC) [45] to generate the super-pixels. Second, we convert the images from red, green, blue (RGB) space to the hue saturation value (HSV) space, and then the color histograms for all the super-pixels, which are 2-D over the H and S channels, are calculated. The super-pixels that overlap with the scribbles are adopted as the foreground (road) samples, and the super-pixels that overlap outside the a_2 buffer are designated as background (nonroad) samples. Accordingly, the cumulative histograms for the foreground and background are calculated. Third, a graph is built where a node represents a super-pixel. For each node, there are two types of corresponding edges. A type of edges connects the node with its neighbor nodes, and the other type of edges connects it with both foreground and background nodes. The energy function is defined as follows:

$$E(x) = \sum_i \psi_i(x_i|\text{Hist}, \text{Sc}) + \sum_{i,j} \psi_{ij}\left(x_i, x_j|\text{Hist}\right). \quad (1)$$

Hist denotes the color histogram for all super-pixels; $\text{Sc} = \{s_r, c_r\}$ denotes the scribble annotation where s_r is the pixels of scribble r and $c_r \in$ {foreground, background} is the category label of scribble r. The unary term ψ_i is formulated as follows:

$$\psi_i(x_i) = \begin{cases} 0, & \text{if } x_i \cap s_r \neq \emptyset \text{ and } c_i = c_r \\ \text{KLDiv}(\text{Hist}_i, \text{Hist}_r), & \text{if } x_i \cap \text{Sc} = \emptyset \\ \infty, & \text{otherwise.} \end{cases} \quad (2)$$

Authorized licensed use limited to: Wuhan University. Downloaded on March 09,2021 at 14:59:38 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

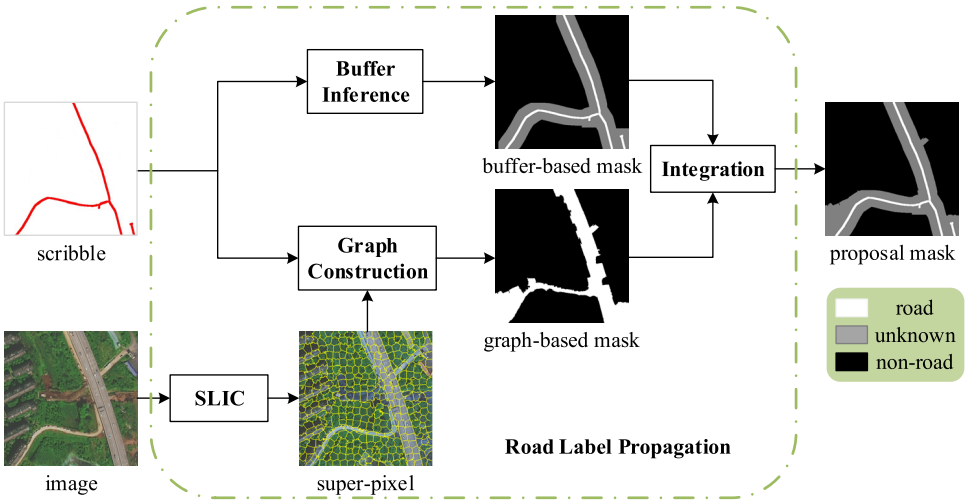


Fig. 2. Process of the road label propagation algorithm.

In the first condition, if super-pixel x_i overlaps with s_r, then it has zero cost when being assigned the label c_r. The second condition means that if x_i does not overlap with any scribbles, then the cost is calculated through Kullback–Leibler divergence (KLDiv) between the normalized histogram of x_i and the cumulative histogram of c_r.

The pairwise term ψ_{ij} evaluates the appearance similarities between two neighbor super-pixels [i.e., x_i and x_j $(x_i \neq x_j)$], by comparing their normalized color histograms using KLDiv. It is assumed that the cost of cutting the edge between two neighbor super-pixels with closer appearance is higher, namely, two neighbor nodes with larger similarities are more likely to have the same label. The graph model propagates label information from the scribbles to the unlabeled pixels to generate a road mask map, but the graph-based masks may contain errors due to the highly varying appearances of the images and the limited capacity of the graph cut-based method.

Finally, we make full use of the buffer inference and graph construction. We integrate the buffer- and graph-based masks based on the following cues: if the pixels denote road in the graph-based mask and nonroad in the buffer-based mask, they are marked as unknown pixels, and the remaining pixels are assigned the same as the buffer-based mask. In this way, we obtain proposal masks which consider not only the buffer-based attributes of road networks but also the color and spatial information obtained from the graph constructed on the super-pixels of the training images. The advantage of our context-aware label propagation algorithm is explicit. The mislabeled pixels from the graph-based masks are hardly distinguished by a standard loss function, which would inevitably impact the results of segmentation; the buffer-based masks assert the absolute discrimination of road and nonroad. In contrast, the labels of unknown pixels of our proposal masks are changeable in a learning-based scheme. The unknown pixels of the proposal masks are classified into potential road or nonroad pixels iteratively through the regularized weakly supervised loss that is described in Section III-B.

B. Dual-Branch Encoder–Decoder Network

As illustrated in Fig. 3, the architecture of our DBNet consists of three parts (one encoder and two decoders) with the RGB channels of an image at a size of 512 × 512 as input. The first part is a feature encoding network using ResNet-34 [16] pretrained on ImageNet [46], and it has five downsampling layers with a minimum scaling ratio of 1/32. Except for the 7 × 7 convolution layer with stride 2 and the first residual blocks with channel number 64, the feature channels are doubled at each downsampling step in the encoder. With regard to the decoding stage, we design two subnetwork branches in parallel: the semantic segmentation branch and the boundary detection branch. Specifically, the segmentation branch uses five transposed convolution layers [47] with stride 2 to restore the resolution of the feature maps from 16 × 16 to 512 × 512, and the feature channels are halved at each upsampling step except for the last two layers. There are three addition skip connections (denoted as circled +) between the encoder feature maps and the decoder (segmentation branch) feature maps. The feature maps with size 256 × 256 in the segmentation branch are concatenated with the corresponding feature maps in the boundary branch. The ASPP module [22], which consists of (a) one 1 × 1 convolution and three parallel 3 × 3 Atrous convolutions with Atrous rates of 1, 2, and 4, respectively, and (b) global average pooling, is applied to the last feature maps of the encoder. The resulting feature maps of ASPP are concatenated and passed through the 1 × 1 convolution layer with channel number 512, and finally fed into the decoder (segmentation branch) part.

For the boundary branch, the multiscale features extracted from the segmentation network are reused. As denoted in Fig. 3, the 512-D features in the decoder (segmentation branch) are first bilinearly upsampled by a factor of 4 and processed by a 3 × 3 convolution layer with channel number 128 and concatenated with the corresponding low-level 128-D features from the encoder, followed by another bilinear upsampling by a factor of 4 and a 3 × 3 convolution layer with channel number 64. The feature maps then are bilinearly upsampled by a factor of 2.

Each convolution layer is activated by the rectified linear unit (ReLU) function except the last convolution layers of the two branches which use sigmoid activation to separately output the probability of each pixel belongs to road surface and boundary.

Authorized licensed use limited to: Wuhan University. Downloaded on March 09,2021 at 14:59:38 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

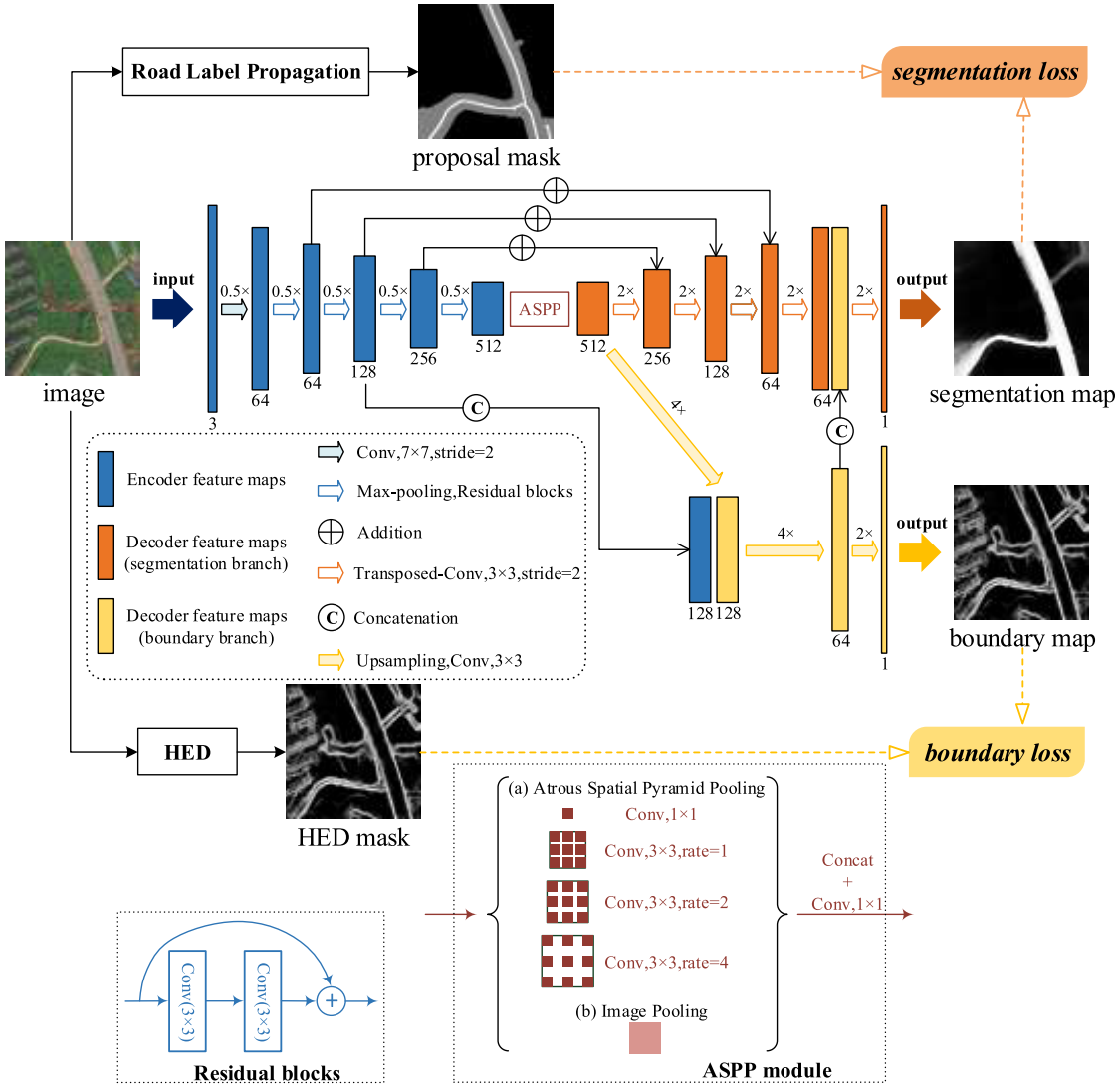


Fig. 3. Structure of the DBNet.

DBNet takes the advantages of encoder–decoder architecture, skip connections, and dual-branch where the boundaries detected by the boundary branch are utilized as prior knowledge to refine and guide the segmentation branch. Here, a holistically nested edge detection (HED) [48] boundary detector pretrained on the generic boundaries of BSDS500 [49] is applied to produce coarse boundary masks, which we called HED masks, without any fine-tuning.

Apart from the HED masks, the proposed network is trained with the proposal masks generated by the road label propagation algorithm described in Section III-A. The segmentation branch and the boundary branch are incorporated under the constraint of a joint loss which combines segmentation loss and boundary loss.

Based on the two categories of labels (known and unknown) provided by proposal mask, the segmentation loss function is

$$L_{\text{seg}} = \text{PBCE}\left(Y_p, S_p\right) + \alpha R(S) \tag{3}$$

where α balances between the partial binary cross-entropy loss (PBCE) and the regularized loss (R) [37].

$\text{PBCE}(Y_p, S_p)$ only computes binary cross-entropy loss between the proposal mask $Y \in \{0.1\}$ and the segmentation map $S \in [0.1]$ for the known pixels $p \in \Omega_k$. $R(S)$ is implemented by the CRF loss with dense Gaussian kernel W over RGBXY channels using fast bilateral filtering [50]

$$\text{PBCE}\left(Y_p, S_p\right) = -\sum_{p\in\Omega_k}\left(Y_p \log S_p + \left(1-Y_p\right)\log\left(1-S_p\right)\right) \tag{4}$$

$$R(S) = S'W(1-S). \tag{5}$$

The gradient of the regularized loss with respect to S is

$$\frac{\partial R(S)}{\partial S} = -2\text{WS}. \tag{6}$$

The loss function of the boundary branch is defined between HED mask T and boundary map B based on the per-pixel mean squared error (MSE) loss, that is

$$L_{\text{bound}} = \frac{1}{w\times h}\sum_{i=1}^{w}\sum_{j=1}^{h}\left(T_{ij}-B_{ij}\right)^2 \tag{7}$$

where w and h are the width and height of the boundary map.

Authorized licensed use limited to: Wuhan University. Downloaded on March 09,2021 at 14:59:38 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

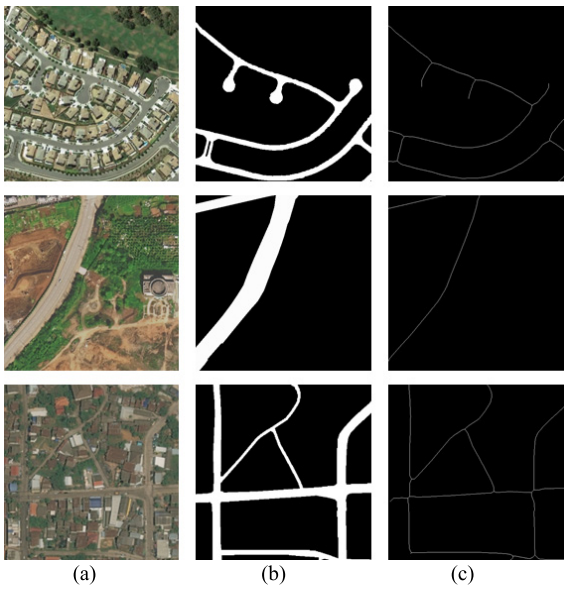


Fig. 4. Examples of the Cheng data set, Wuhan data set, and DeepGlobe data set. (a) Image. (b) Road surface ground truth. (c) Road centerline ground truth.

Overall, the joint loss is

$$L = L_{\text{seg}} + \beta L_{\text{bound}} \tag{8}$$

where β is a coefficient to balance the segmentation loss L_{seg} and the boundary loss L_{bound}.

IV. Experiment and Analysis

A. Data Sets and Evaluation Metrics

We performed our experiments on three diverse road data sets: 1) the Cheng data set [51]; 2) the Wuhan data set; and 3) the DeepGlobe data set [52]. These data sets are comprised of high-resolution aerial and satellite images from urban, suburban, and rural regions covering a total area of approximately 1665 km^2 across the world with varied ground sampling distance (GSD) from 0.5 to 1.2 m. All the images and the corresponding ground truths were seamlessly cropped into 512 × 512 tiles, and then divided into training set and test set. For the Cheng data set, we followed the splitting rules of [51]; for the Wuhan data set and the DeepGlobe data set, the ratio between training and testing samples is 3:1. The details of the datasets are listed in Table I, and examples are shown in Fig. 4. In our experiments, the road centerline was utilized as a typical scribble supervision. Specifically, the ground truth of the Cheng data set included manually labeled road surface and centerline, whereas there was only pixel-wise annotated road surface ground truth for the Wuhan data set and the DeepGlobe data set; therefore, we skeletonized the road surface ground truth to obtain the centerline ground truth for the last two road data sets.

The precision, recall, F_1 score, and intersection over union (IoU), which were widely used as indicators in the related literature on road segmentation (see [12], [20], [21]), were adopted to evaluate the segmentation accuracy of the road extraction results at the pixel level. The precision was the fraction of the predicted road pixels that were true roads, and the recall was the fraction of the true road pixels that were correctly predicted. The F_1 and IoU were the overall metrics that offered a tradeoff between precision and recall. More precisely, the IoU was the ratio between the intersection of the predicted road pixels and the labeled road pixels and the results of their union. These pixel-level evaluation metrics are defined as follows:

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

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

$$F_1 = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{11}$$

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

where TP, FP, and FN represent the true positive, false positive, and false negative, respectively.

TABLE I
Details of the Experimental Road Data Sets

Dataset	Area (km^2)	Source	GSD(m)	Tiles(train/test)
Cheng [51]	132	aerial	1.2	300 / 49
Wuhan	200	satellite	0.5	1944 / 648
DeepGlobe [52]	1333	satellite	0.5	15000 / 5333

B. Implementation Details

The road label propagation algorithm integrates the buffer-based mask with the graph-based mask to generate the proposal mask. For the buffer inference, a_1 was set smaller than the minimum road width and a_2 was set close to the maximum road width. For example, we set $a_1 = 6$ m and $a_2 = 18$ m for the Cheng data set with road widths ranging from 12 to 18 m. We set $a_1 = 2$ m for the Wuhan data set and the DeepGlobe data set, and set $a_2 = 38$ m for the Wuhan data set and 9 m for the DeepGlobe data set, respectively. The impact of different buffer widths is further analyzed in Section IV-D. With respect to the graph construction, the SLIC super-pixels were calculated, the approximate number of which was 400 on 512 × 512 image patch, and the compactness was set as 20 to balance the color-space proximity. To find neighbors of the super-pixels, we utilized Delaunay tessellation for simplicity. The color histograms for all the super-pixels were built on the HSV space using 20 bins for the H and S channels; all the bins were concatenated and normalized; and the ranges of H and S were [0, 360] and [0, 1], respectively.

Before training the DBNet, we employed the HED boundary detector which was pretrained on the generic boundaries of BSDS500 [49] to predict HED masks. All the HED masks were seamlessly cropped into 512 × 512 tiles as well. We implemented data augmentation, including image horizontal flip, vertical flip, diagonal flip, color jittering, shifting, and scaling. In terms of the regularized loss, the Gaussian bandwidths for RGB (color domain) and XY (spatial domain) were set as 15 and 100, respectively. The loss weight α of the regularized loss was set to 0.5 in (3); and the loss weight β, which balances between the segmentation loss and the boundary loss, was set at 0.7. During the training phase,

Authorized licensed use limited to: Wuhan University. Downloaded on March 09,2021 at 14:59:38 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE II
EXPERIMENTAL RESULTS OF DIFFERENT WEAKLY SUPERVISED LEARNING METHODS ON THREE ROAD DATA SETS, WHERE THE VALUES IN BOLD ARE WITH THE BEST PERFORMANCE

Dataset	Method	Precision	Recall	F_1	IoU
Cheng	ScribbleSup [25]	0.5274	**0.9730**	0.6750	0.5190
	BPG [39]	0.7179	0.9085	0.7925	0.6627
	WSOD [40]	0.8468	0.8852	0.8608	0.7594
	WeaklyOSM [44]	0.7798	0.9077	0.8322	0.7170
	ScRoadExtractor (Ours)	**0.9033**	0.8423	**0.8657**	**0.7651**
Wuhan	ScribbleSup [25]	0.6086	**0.7267**	0.6222	0.4740
	BPG [39]	**0.7510**	0.5804	0.6197	0.4717
	WSOD [40]	0.7143	0.5329	0.5789	0.4298
	WeaklyOSM [44]	0.7509	0.6020	0.6300	0.4805
	ScRoadExtractor (Ours)	0.6963	0.6904	**0.6580**	**0.5158**
DeepGlobe	ScribbleSup [25]	0.2951	**0.8813**	0.4079	0.2694
	BPG [39]	0.6681	0.7638	0.6624	0.5157
	WSOD [40]	0.6549	0.6265	0.5899	0.4438
	WeaklyOSM [44]	0.7115	0.7378	0.6673	0.5239
	ScRoadExtractor (Ours)	**0.7954**	0.7138	**0.7132**	**0.5782**

we set the total epochs as 300, and it terminated earlier if the total loss stopped decreasing over six continuous epochs. The Adam optimizer [53] was selected as the network optimizer. The batch size was fixed as two on the 512×512 tiles. The learning rate was initially set at $2e^{-4}$, and divided by 5 if the total loss stopped decreasing up to three continuous epochs. The implementation is based on PyTorch. On a single NVIDIA GTX1060 GPU with 6-GB memory, the training process of DBNet took about 0.5, 7.5, and 23.8 h for the Cheng data set, the Wuhan data set, and the DeepGlobe data set, respectively.

It should be noted that the road label propagation (i.e., buffer inference and graph construction) and HED masks generation were only used for training and not needed for testing. Therefore, in the testing phase, we only applied the DBNet on the test images. The test time augmentation (TTA) was adopted, which included image horizontal flip, vertical flip, and diagonal flip (predicting each $2 \times 2 = 8$ times) also on 512×512 tiles. The output probability of each pixel from the sigmoid classifier was translated to binary values with a threshold of 0.5. The road surface extraction results (i.e., the segmentation map derived from the segmentation branch) were evaluated using the four aforementioned metrics.

C. Experimental Results

We evaluated our proposed ScRoadExtractor on the Cheng data set, Wuhan data set, and DeepGlobe data set and compared its performance with recent scribble-based weakly supervised segmentation methods, including the classic ScribbleSup [25], which adopted an alternative training scheme between proposal generation and network training; boundary perception guidance (BPG) [39], which combined scribbles and rough edge maps for supervision to guide the segmentation network; a weakly supervised salient object detection (WSOD) method [40]; and a method specially for weakly supervised road segmentation using OSM, named WeaklyOSM [44] in this article.

As shown in Table II, ScRoadExtractor achieved the best results in both F_1 and IoU compared with the other approaches on all the data sets. For ScribbleSup, the alternation between proposal generation and network training happens when training converges; and in this article, we show its results after three alternations in Table II. Obviously, ScribbleSup performed poorly on these road data sets, and it was computationally expensive due to the alternative training scheme. In contrast, even a single round of training was enough to improve for the end-to-end methods (e.g., BPG, WSOD, WeaklyOSM, ScRoadExtractor). In terms of the Cheng data set, ScRoadExtractor outperformed BPG by 7.32% in F_1 and 10.24% in IoU. Similarly, on the Wuhan data set and the DeepGlobe data set, the IoU of ScRoadExtractor was 4.41% and 6.25% higher than BPG, respectively. The BPG had the weakness that the semantics and boundaries interact only at the loss functions without considering the correlation of the features between two subnetworks. ScRoadExtractor performed slightly better than WSOD on the smallest Cheng data set but significantly better on the Wuhan data set and DeepGlobe data set, which indicated WSOD only handled with relatively simple scenarios and lacked of strong generalization ability. Taking the results of WeaklyOSM (the second best) as a baseline, ScRoadExtractor achieved 4.81%, 3.53%, and 5.43% growth in IoU, on the Cheng data set, Wuhan data set, and DeepGlobe data set, respectively.

Figs. 5–7 show some examples of the road segmentation results predicted by different methods on the 512×512 tiles selected from these road data sets. Please note that the scribble annotation (b) is not required for testing. It can be seen that the results of ScribbleSup (c) contain many nonroad pixels with poor boundary localization. As illustrated in the first rows of Figs. 6 and 7, BPG (d), WSOD (e) and WeaklyOSM (f) have faced difficulty in correctly identifying the roads shaded by buildings and trees from satellite images, resulting in missing road segments and ambiguous boundaries; but ScRoadExtractor (g) was robust to occlusions and shadows. Furthermore, it can be seen that ScRoadExtractor (g) achieved a segmentation map most similar to the per-pixel annotated road surface ground truth (h) with better boundary alignment, which demonstrates that ScRoadExtractor can extract the road surface more reliably from remote sensing images.

Authorized licensed use limited to: Wuhan University. Downloaded on March 09,2021 at 14:59:38 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

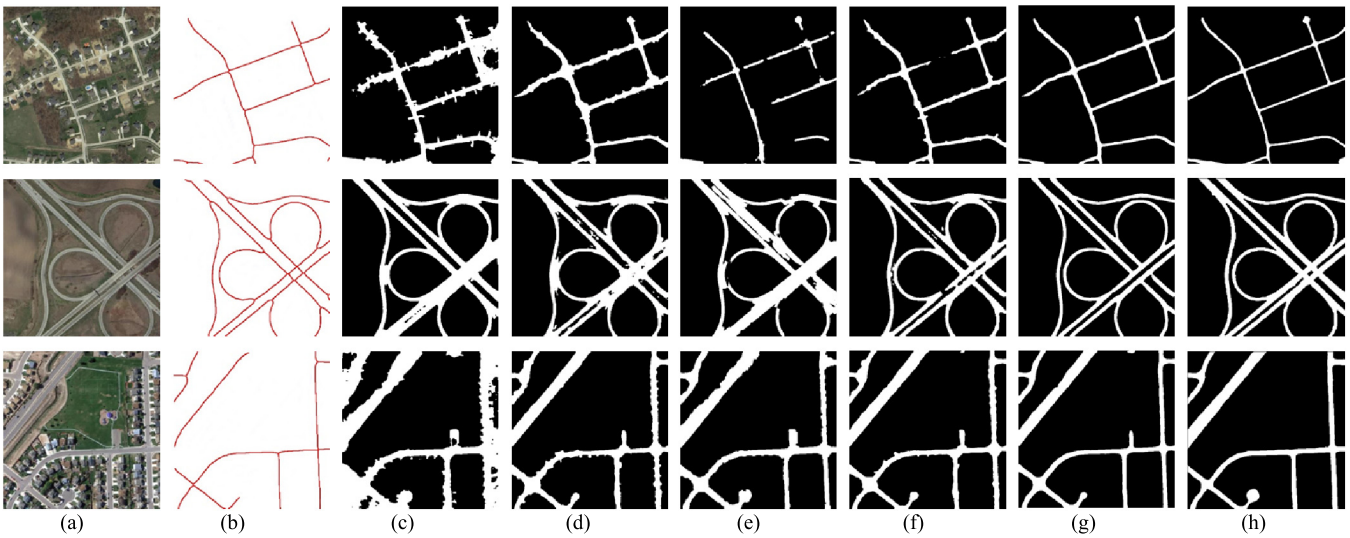


Fig. 5. Qualitative results of road segmentation using different methods on the Cheng data set. (a) Image. (b) Scribble annotation. (c) ScribbleSup. (d) BPG. (e) WSOD. (f) WeaklyOSM. (g) ScRoadExtractor. (h) Per-pixel annotation (ground truth).

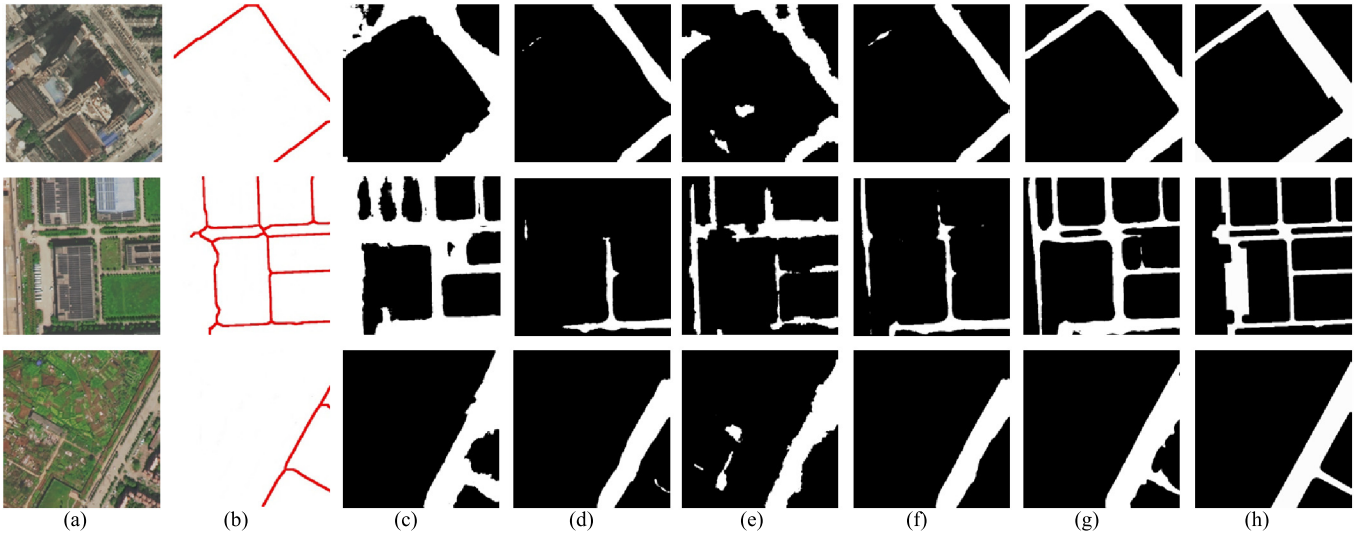


Fig. 6. Qualitative results of road segmentation using different methods on the Wuhan data set. (a) Image. (b) Scribble annotation. (c) ScribbleSup. (d) BPG. (e) WSOD. (f) WeaklyOSM. (g) ScRoadExtractor. (h) Per-pixel annotation (ground truth).

WeaklyOSM presented in [44] has a problem formulation similar to our study, but ScRoadExtractor differs from it in the following aspects. First, the initial road annotation generation of WeaklyOSM, which had only a buffer-based road width inference, relied on the fixed scribble annotation (i.e., centerline), while other general forms of scribble annotations can be applied by the road label propagation algorithm of ScRoadExtractor. Second, an auxiliary boundary branch was newly employed by ScRoadExtractor to refine and enhance the performance of weakly supervised semantic segmentation. The function of the boundary branch can be seen from Figs. 5–7 that ScRoadExtractor had much better road connectivity through learning from continuous boundaries, and the corresponding ablation study of Section IV-D. To demonstrate the generalization ability of ScRoadExtractor, we introduced the simulated scribbles on the Wuhan data set that were created by eroding the road surface ground truth with a cross-shaped kernel of size 7 and offset anchor (3, 6), followed by the skeletonization. Table III shows that our method exceeded WeaklyOSM by 6.44% on F_1 and 6.94% on IoU under the same supervision of the simulated scribbles on the Wuhan data set. In contrast to [44], ScRoadExtractor generalized well on different forms of scribble annotations, ranging from road centerline ground truth to simulated scribbles, without the limitation of the hard constraint at the early stages.

D. Ablation Study

In this section, the impact of road label propagation algorithm was analyzed by using different buffer widths and different supervision strategies to train DBNet. In addition, an ablation study was performed to verify the effectiveness of

Authorized licensed use limited to: Wuhan University. Downloaded on March 09,2021 at 14:59:38 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

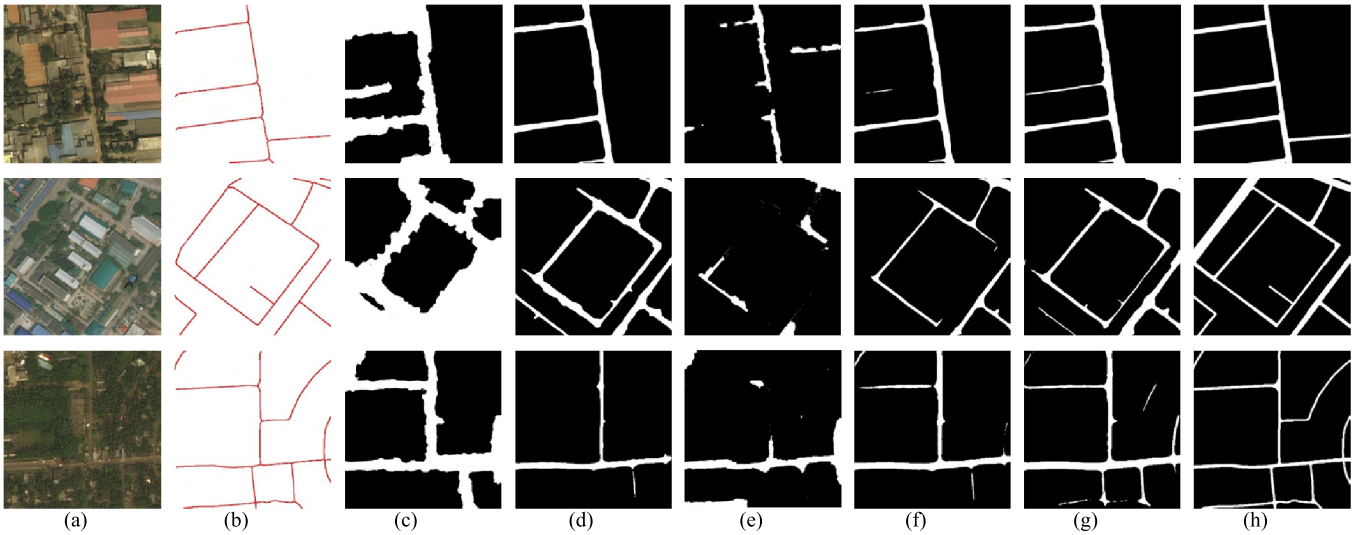


Fig. 7. Qualitative results of road segmentation using different methods on the DeepGlobe data set. (a) Image. (b) Scribble annotation. (c) ScribbleSup. (d) BPG. (e) WSOD. (f) WeaklyOSM. (g) ScRoadExtractor. (h) Per-pixel annotation (ground truth).

TABLE III
COMPARISON RESULTS BASED ON SIMULATED SCRIBBLES ON THE WUHAN DATA SET

Method	Precision	Recall	F_1	IoU
WeaklyOSM [44]	0.8661	0.4687	0.5759	0.4307
ScRoadExtractor (Ours)	0.8318	0.5680	0.6403	0.5001

the proposed boundary branch, and we compared it with BRN presented in [39].

First, we evaluated the impact of different buffer widths in proposal mask generation. Parameter a_1 defined the scope of pure road pixels. It could be well estimated according to the minimum road width. Specifically, a_1 was set at 6 m for the Cheng data set, and 2 m for the Wuhan data set and the DeepGlobe data set as the latter two cover suburban and rural areas with many narrow roads. Here, we focused on the impact of parameter a_2, which was a tradeoff between buffer inference and graph construction. The results of using different a_2 values were shown in Fig. 8. It was observed that the road surface extraction got the best performance when a_2 was set at the maximum road width, e.g., 18 m for the Cheng data set. When a_2 was increased or decreased gradually, the performance got worse. If a_2 was too large, the proposal mask approached to the empirical buffer-based mask but in fact the road widths varied; if a_2 was too small, the proposal mask was closer to the graph-based mask which may contain noise due to the limited capacity of graph cut method. Finally, a_2 was set at 18, 38, and 9 m for the Cheng data set, the Wuhan data set, and the DeepGlobe data set, respectively.

Second, we investigated the effects of weak supervision and full supervision on training the DBNet with different labels. The "expanded mask" in Table IV indicated directly expanding the centerline with a certain road width (e.g., 10 m); the buffer-based mask and the graph-based mask were intermediate products of our proposed road label propagation algorithm (see Fig. 2). Here and in the next experiment, a_2 was set at 15 m

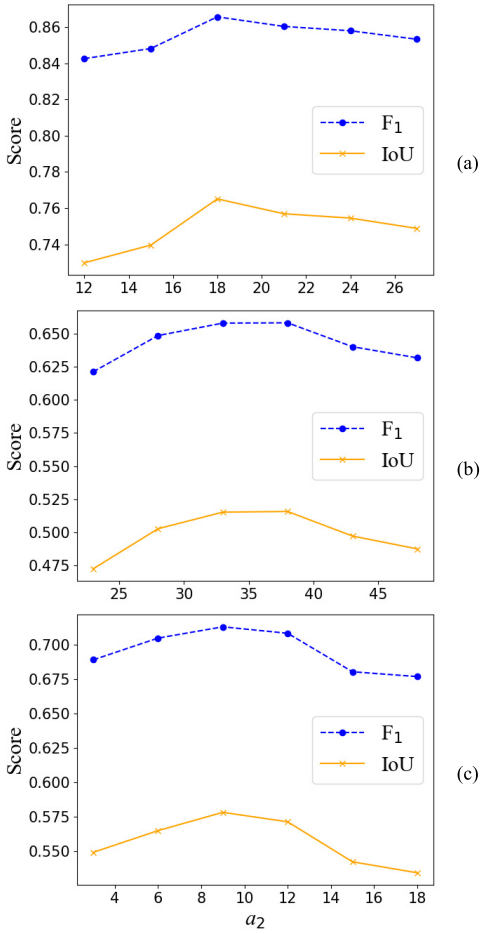


Fig. 8. F_1 (blue dotted line) and IoU (yellow solid line) curves for different buffer widths a_2 on the Cheng data set (a), the Wuhan data set (b), and the DeepGlobe data set (c).

for the Cheng and the DeepGlobe data set, and 29 m for the Wuhan data set. As can be seen from Table IV, DBNet trained with our proposed proposal mask obtained the best results

Authorized licensed use limited to: Wuhan University. Downloaded on March 09,2021 at 14:59:38 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE IV
ABLATION STUDY ABOUT DIFFERENT SUPERVISIONS FOR ROAD SEGMENTATION ON THREE ROAD DATA SETS

Supervision	Label	Cheng		Wuhan		DeepGlobe	
		F_1	IoU	F_1	IoU	F_1	IoU
Weak	graph-based mask	0.6555	0.4972	0.6165	0.4689	0.3980	0.2614
	expand mask	0.8280	0.7089	0.5467	0.3912	0.6203	0.4678
	buffer-based mask	0.8033	0.6769	0.6325	0.4848	0.6666	0.5224
	proposal mask	0.8482	0.7396	0.6484	0.5028	0.6805	0.5422
Full	full mask	0.9239	0.8597	0.6956	0.5648	0.7617	0.6408

TABLE V
IMPACT OF BOUNDARY BRANCH

Backbone (Encoder)	Branch (Decoder)	Parameters (M)	FLOPs (G)	Loss			Cheng		Wuhan		DeepGlobe	
				PBCE	*R*	L_{bound}	F_1	IoU	F_1	IoU	F_1	IoU
ResNet-34	w/ *Seg*	30.5759	33.3567	√			0.7987	0.6729	0.6404	0.4944	0.6502	0.5070
	w/ *Seg* + BRN	30.5765	33.3568	√		√	0.7634	0.6233	0.6422	0.4948	0.6712	0.5269
	w/ *Seg* + *Bou*	31.3217	45.9991	√	√	√	0.8482	0.7396	0.6484	0.5028	0.6805	0.5422

with respect to other weak supervision strategies and achieved acceptable results compared with the per-pixel annotated full mask supervision.

Third, we explored the impact of the auxiliary boundary branch (shorted as Bou). The comparison results of different branches based on the ResNet-34 backbone were shown in Table V, which were all trained with the same proposal masks. The BRN presented in [39] and our boundary branch (Bou) were trained and optimized with the same HED masks. For the Cheng data set, with the segmentation branch (Seg) alone, the model obtained 79.87% in F_1. Adding BRN, F_1 decreased by 3.53%, whereas F_1 achieved 4.41% improvement by combining our Bou. In terms of the Wuhan data set, the best results were produced by our proposed DBNet (i.e., w/Seg+Bou), although these network structures performed similarly. Compared with employing Seg only and introducing BRN, DBNet improved 3.52% and 1.53% in IoU on the DeepGlobe data set, respectively. The results demonstrated the generalization ability and effectiveness of our DBNet with the well-designed boundary branch. There are two critical differences between BRN and our proposed boundary branch. First, only one-shot upsampling layer was used in BRN, while in Bou (see Fig. 3), multiple upsampling layers were employed as well as feature sharing with Seg. Second, BRN lacked of graph-based regularizations to capture global and local dependencies between known (road and nonroad) and unknown pixels. Both of which leaded to it worse performance than ours and even the backbone (w/Seg) on the Cheng data set.

V. DISCUSSION

Deep learning has made remarkable achievements in many research subjects, especially vision-based tasks. At the same time, the requirement of huge training data sets is also criticized and rethought. Although the trend of collecting vast amounts of data for feeding deep networks is still on-going, discovering knowledge with less training data (few-shot learning or weakly supervised learning) or without training data (unsupervised learning) has drawn increasingly more attention and begun to form another mainstream. In remote sensing image processing, the lack of high-quality and up-to-date ground truths makes the deep learning-based approaches hardly applicable to new remote sensing images. Integrating city-scale or larger scale open-source maps, such as VGI, and the burgeoning weakly supervised learning approach to reduce the demand of training data are a promising area of research as well.

Our proposed ScRoadExtractor is an ideal instance of weakly supervised learning that only utilizes sparse scribble annotations instead of densely annotated ground truths for road surface segmentation. However, this difference between our method and recent other weakly supervised methods is distinct. The core of ScRoadExtractor is a road label propagation algorithm, which generates the proposal masks from scribbles by aggregating the buffer-based properties of roads and the continuity of similar features in the space and color domains, to mark each pixel as known (i.e., road or nonroad) or unknown. Compared with the buffer-based masks (as [44]) and the graph-based masks (as [25]), our proposal masks benefit from a better balance between the utilizations of foreground and background information. Based on the proposal masks, we designed a DBNet containing a semantic segmentation branch and a boundary detection branch, which interacted both at the feature level and the output level, while [39] associated them only with the loss functions.

The scribble annotations used in ScRoadExtractor are not restricted to road centerlines from a GIS map or OSM data. Different from [43] and [44], which assume perfect centerlines, more scribble forms can be utilized in ScRoadExtractor; and a commonly used candidate is GPS traces from vehicles or pedestrians. This kind of scribble can be widely accessed from many open-source databases or websites. Although they are not that accurate, ScRoadExtractor can process them with the combination of buffer- and graph-based mask generation and boundary alignment. The provided road surface information can in turn aid GPS for better traffic management and navigation.

A point should be mentioned is the edge detector we adopted. In this work, we used HED pretrained on the open data set BSDS500 [49] as the edge detector. In fact, we also tested classic edge detectors such as Canny [54] and Sobel [55]. We found that the effect of using these detectors was also satisfactory and only slightly worse than using HED.

Authorized licensed use limited to: Wuhan University. Downloaded on March 09,2021 at 14:59:38 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

For example, in the ablation study of Wuhan data set, the IoU was 0.493 when using Canny and was 0.503 when using HED. They can be alternatives to HED.

Our method outperformed the state-of-the-art scribble-based weakly supervised segmentation methods as shown in Table IV; nevertheless, it has not fully filled the gap between weak supervision and full supervision as their performance difference is still noticeable. There is a lot of work to be done in future research, which should include the following. First, the graph construction in the road label propagation step can be further formulated as graph representation learning (e.g., via a graph convolution network (GCN) [56]), which can embed the topological information of road networks into the learning-based graph structure. Second, the boundary regularization of road networks will be an important step toward the level of manual delineation, for example, a rotation Gaussian-Mask [57] may be designed to model a road segment and to solve the boundary mislabeling problem. Third, with access to a small number of full (pixel-level) annotations and a large number of weak (scribble) annotations, the proposed method may be able to match the performance of full supervision with semi-supervised learning methods, e.g., an adversarial self-taught learning framework [58] for semi-supervised semantic segmentation.

VI. Conclusion

In this article, we proposed a scribble-based weakly supervised learning method, called ScRoadExtractor, for road surface segmentation from remote sensing images, which employs an end-to-end training scheme that can achieve good results without the need for alternating optimization. To propagate semantic information from scribble annotations to unlabeled pixels, we introduced a new road label propagation algorithm to generate proposal masks, which integrate the buffer-based masks inferred from the buffer-based strategy and the graph-based masks obtained from the graph constructed on the super-pixels. In addition, we introduced a DBNet, in which we inject the boundary information into semantic segmentation and also a joint loss function that refines both the semantic and boundary predictions. Taking the road centerline as a typical form of scribble annotations in our experiments, we showed that our method was superior to recent related methods and further demonstrated that ScRoadExtractor can be generalized to general forms of scribble annotations.

At present, very few works have explored weakly supervised semantic segmentation for extracting road surfaces from remote sensing images. Our method is a step on a journey that will ultimately bring us closer to automatic road extraction from remote sensing images with very little manual annotating required. We further believe that although ScRoadExtractor was originally designed for road segmentation, we anticipate that it also can be adapted to other segmentation tasks.

References

[1] V. Mnih and G. E. Hinton, "Learning to detect roads in high-resolution aerial images," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 210–223.

[2] J. D. Wegner, J. A. Montoya-Zegarra, and K. Schindler, "A higher-order CRF model for road network extraction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1698–1705.

[3] R. Alshehhi and P. R. Marpu, "Hierarchical graph-based segmentation for extracting road networks from high-resolution satellite images," *ISPRS J. Photogramm. Remote Sens.*, vol. 126, pp. 245–260, Apr. 2017.

[4] C. Tao, J. Qi, Y. Li, H. Wang, and H. Li, "Spatial information inference net: Road extraction using road-specific contextual information," *ISPRS J. Photogramm. Remote Sens.*, vol. 158, pp. 155–166, Dec. 2019.

[5] X. Zhang, X. Han, C. Li, X. Tang, H. Zhou, and L. Jiao, "Aerial image road extraction based on an improved generative adversarial network," *Remote Sens.*, vol. 11, no. 8, p. 930, Apr. 2019.

[6] X. Huang and L. Zhang, "Road centreline extraction from high-resolution imagery based on multiscale structural features and support vector machines," *Int. J. Remote Sens.*, vol. 30, no. 8, pp. 1977–1987, Apr. 2009.

[7] Z. Miao, W. Shi, H. Zhang, and X. Wang, "Road centerline extraction from high-resolution imagery based on shape features and multivariate adaptive regression splines," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 3, pp. 583–587, May 2013.

[8] G. Cheng, F. Zhu, S. Xiang, Y. Wang, and C. Pan, "Accurate urban road centerline extraction from VHR imagery via multiscale segmentation and tensor voting," *Neurocomputing*, vol. 205, pp. 407–420, Sep. 2016.

[9] L. Gao, W. Shi, Z. Miao, and Z. Lv, "Method based on edge constraint and fast marching for road centerline extraction from very high-resolution remote sensing images," *Remote Sens.*, vol. 10, no. 6, p. 900, Jun. 2018.

[10] F. Bastani *et al.*, "RoadTracer: Automatic extraction of road networks from aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4720–4728.

[11] M. Haklay and P. Weber, "OpenStreetMap: User-generated street maps," *IEEE Pervas. Comput.*, vol. 7, no. 4, pp. 12–18, Oct. 2008.

[12] T. Panboonyuen, K. Jitkajornwanich, S. Lawawirojwong, P. Srestasathiern, and P. Vateekul, "Road segmentation of remotely-sensed images using deep convolutional neural networks with landscape metrics and conditional random fields," *Remote Sens.*, vol. 9, no. 7, p. 680, Jul. 2017.

[13] 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, Dec. 2017.

[14] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[15] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.

[18] Y. Xu, Z. Xie, Y. Feng, and Z. Chen, "Road extraction from high-resolution remote sensing imagery using deep learning," *Remote Sens.*, vol. 10, no. 9, p. 1461, Sep. 2018.

[19] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[20] L. Zhou, C. Zhang, and M. Wu, "D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 192–196.

[21] H. He, D. Yang, S. Wang, S. Wang, and Y. Li, "Road extraction by using atrous spatial pyramid pooling integrated encoder-decoder network and structural similarity loss," *Remote Sens.*, vol. 11, no. 9, p. 1015, Apr. 2019.

[22] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*. [Online]. Available: http://arxiv.org/abs/1706.05587

[23] X. Zhang, W. Ma, C. Li, J. Wu, X. Tang, and L. Jiao, "Fully convolutional network-based ensemble method for road extraction from aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 10, pp. 1777–1781, Oct. 2020.

[24] Y. Wei, K. Zhang, and S. Ji, "Simultaneous road surface and centerline extraction from large-scale remote sensing images using CNN-based segmentation and tracing," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8919–8931, Dec. 2020.

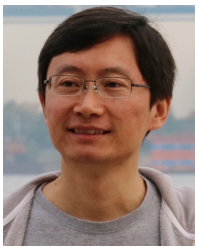
Authorized licensed use limited to: Wuhan University. Downloaded on March 09,2021 at 14:59:38 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

[25] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "ScribbleSup: Scribble-supervised convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3159–3167.
[26] P. Vernaza and M. Chandraker, "Learning random-walk label propagation for weakly-supervised semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2953–2961.
[27] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei, "What's the point: Semantic segmentation with point supervision," 2015, *arXiv:1506.02106*. [Online]. Available: http://arxiv.org/abs/1506.02106
[28] J. Dai, K. He, and J. Sun, "BoxSup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," 2015, *arXiv:1503.01640*. [Online]. Available: http://arxiv.org/abs/1503.01640
[29] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele, "Simple does it: Weakly supervised instance and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 876–885.
[30] J. Xu, A. G. Schwing, and R. Urtasun, "Learning to segment under various forms of weak supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3781–3790.
[31] J. Ahn and S. Kwak, "Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4981–4990.
[32] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, Nov. 2001.
[33] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1124–1137, Sep. 2004.
[34] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille, "Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1742–1750.
[35] A. Kolesnikov and C. H. Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 695–711.
[36] M. Tang, A. Djelouah, F. Perazzi, Y. Boykov, and C. Schroers, "Normalized cut loss for weakly-supervised CNN segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1818–1827.
[37] M. Tang, F. Perazzi, A. Djelouah, I. B. Ayed, C. Schroers, and Y. Boykov, "On regularized losses for weakly-supervised CNN segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 524–540.
[38] A. Obukhov, S. Georgoulis, D. Dai, and L. Van Gool, "Gated CRF loss for weakly supervised semantic image segmentation," 2019, *arXiv:1906.04651*. [Online]. Available: http://arxiv.org/abs/1906.04651
[39] B. Wang *et al.*, "Boundary perception guidance: A scribble-supervised semantic segmentation approach," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 3663–3669.
[40] J. Zhang, X. Yu, A. Li, P. Song, B. Liu, and Y. Dai, "Weakly-supervised salient object detection via scribble annotations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12546–12555.
[41] J. Yuan and A. M. Cheriyadat, "Road segmentation in aerial images by exploiting road vector data," in *Proc. 4th Int. Conf. Comput. Geospatial Res. Appl.*, Jul. 2013, pp. 16–23.
[42] G. Mattyus, S. Wang, S. Fidler, and R. Urtasun, "Enhancing road maps by parsing aerial images around the world," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1689–1697.
[43] 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, Nov. 2017.
[44] 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, p. 478, Oct. 2019.
[45] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.
[46] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Apr. 2015.
[47] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2018–2025.
[48] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1395–1403.
[49] 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, May 2011.
[50] A. Adams, J. Baek, and M. A. Davis, "Fast high-dimensional filtering using the permutohedral lattice," *Eurographics*, vol. 29, no. 2, pp. 753–762, May 2010.
[51] G. Cheng, Y. Wang, S. Xu, H. Wang, S. Xiang, and C. Pan, "Automatic road detection and centerline extraction via cascaded end-to-end convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 6, pp. 3322–3337, Jun. 2017.
[52] I. Demir *et al.*, "DeepGlobe 2018: A challenge to parse the Earth through satellite images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 172–181.
[53] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[54] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986.
[55] I. E. Sobel, "Camera models and machine perception," Doctoral Dissertation, Stanford Univ., Stanford, CA, USA, 1970.
[56] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*. [Online]. Available: http://arxiv.org/abs/1609.02907
[57] X. Zhang, G. Wang, P. Zhu, T. Zhang, C. Li, and L. Jiao, "GRS-Det: An anchor-free rotation ship detector based on Gaussian-mask in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, early access, Sep. 4, 2020, doi: 10.1109/TGRS.2020.3018106.
[58] W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang, "Adversarial learning for semi-supervised semantic segmentation," 2018, *arXiv:1802.07934*. [Online]. Available: http://arxiv.org/abs/1802.07934

Yao Wei received the B.S. degree in geographic information science from the China University of Petroleum, Qingdao, China, in 2018. She is pursuing the M.S. degree in photogrammetry and remote sensing with Wuhan University, Wuhan, China.

Her research interests include deep learning and remote sensing image analysis.



Shunping Ji (Member, IEEE) received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2007.

He is a Professor with the School of Remote Sensing and Information Engineering, Wuhan University. He has coauthored more than 50 articles. His research interests include photogrammetry, remote sensing image processing, mobile mapping system, and machine learning.

Authorized licensed use limited to: Wuhan University. Downloaded on March 09,2021 at 14:59:38 UTC from IEEE Xplore. Restrictions apply.