Road Extraction From Remote Sensing Images in Wildland–Urban Interface Areas

Ruonan Chen, Xiang Li, Yuan Hu, Congcong Wen, *Student Member, IEEE*, and Ling Peng

***Abstract*— In this letter, we address the problem of road extraction in Wildland–urban interface (WUI) areas. In recent years, with the great success of convolutional neural networks (CNNs) in various vision-related tasks, researchers have developed many CNN-based methods for road extraction on remote sensing images. Nevertheless, these methods mostly treat road extraction as a binary classification problem on semantic labeling. In WUI areas, the road is narrower and tends to be occluded by trees, which may result in the serious discontinuous problem of inferred road maps. To address this issue, we propose transforming the input representation of the binary classification map into a continuous signed distance map. In this way, our model is forced to predict the continuous distance representations and, thus, improve the spatial continuities of inferred roads. In addition, a real-value regression task is designed to train along with the original binary classification task to generate spatially continuous and semantically accurate road maps. Then, we conduct experiments on the public Massachusetts road data set and a homemade data set collected from Yajishan Mountain, Beijing, China. Finally, our proposed method achieves intersection-over-unions (IoUs) of 64.11% and 65.92% for the Massachusetts and WUI-Yajishan data sets, respectively, without any postprocessing. In addition, the ablation analysis shows that introducing the regression task on the proposed signed distance representation can effectively alleviate the problem of discontinuous road prediction. Furthermore, comparing with the state-of-the-art methods demonstrates the superiority of our method for road extraction in WUI areas.**

***Index Terms*— Convolutional neural networks (CNNs), remote sensing image, road extraction, signed distance representation.**

I. Introduction

ROAD information is one of the most important geographic information elements and plays an important role in various applications, such as urban planning, autonomous driving, and emergency rescue. Although the substantial focus has been on automatic road extraction from remote sensing images, current studies mostly focus on road extraction in urban areas. Road extraction in Wildland–urban interface (WUI) areas has been less studied, but it is meaningful for emergency rescue. Moreover, unlike urban areas, the roads in WUI areas are narrower and tend to be occluded by trees, which makes the automatic road extraction problem more challenging. In this letter, we address the problem of automatic road extraction from remote sensing images in WUI areas.

Before the era of deep learning, traditionally shallow models mostly relied on the handcrafted design of road features, such as linear shape, color, gradient, and energy. For example, Marikhu *et al.* [1] used the cooperating snakes method with an energy function, designed by gradient vector flow. Hu *et al.* [2] first detected road intersections based on shape features and then used a toe-finding algorithm to find the dominant directions to track a road segment. Song and Civco [3] adopted the shape index feature and support vector machine (SVM) to detect roads. Singh and Garg [4] combined adaptive global thresholding and morphological operations to extract roads. In summary, these traditional methods usually need expert knowledge for the feature-engineering process and have low generalization abilities when dealing with new data sets, which limits their applications in real-world scenarios.

Recently, deep learning-based methods, especially deep convolutional neural networks (CNNs), have achieved remarkable performance in various vision-related tasks, as well as in remote sensing image analysis. Following this trend, recent studies have mostly adopted deep CNN models for automatic road extraction from remote sensing images. For example, the work in [5] was one of the earliest attempts to design an effective deep neural network for automatic road extraction from aerial images. In this method, the dense classification maps are generated by using a sliding window over the input images, and the model predicts one semantic label for each image patch at one time. Following methods [6], [7] also perform semantic labeling in a patch-based manner.

Among all deep learning-based semantic segmentation models, fully convolutional networks (FCNs) represent a milestone for dense semantic labeling. In the FCN model, the fully connected layers are replaced by convolutional layers and, therefore, enable semantic segmentation without using sliding patches. After that, various FCN-like architectures have been proposed to improve the performance of semantic segmentation; notable methods include U-Net [8], SegNet [9], and DeepLabV3+ [10]. In this direction, Zhang *et al.* [11] proposed a deep residual U-Net for road extraction and demonstrated superior performance to patch-based methods on the Massachusetts data set. Xin *et al.* [12] proposed a road

Manuscript received July 29, 2020; revised September 10, 2020; accepted September 26, 2020. Date of publication October 16, 2020; date of current version December 16, 2021. This work was supported by the Beijing Municipal Science and Technology Project under Grant Z191100001419002. *(Ruonan Chen and Xiang Li contributed equally to this work.) (Corresponding author: Ling Peng.)*

Ruonan Chen, Yuan Hu, and Congcong Wen are with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China, and also with the College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China.

Xiang Li is with the Department of Electrical and Computer Engineering, NYU Abu Dhabi, Abu Dhabi 25586, UAE.

Ling Peng is with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China (e-mail: pengling@aircas.ac.cn).

Digital Object Identifier 10.1109/LGRS.2020.3028468

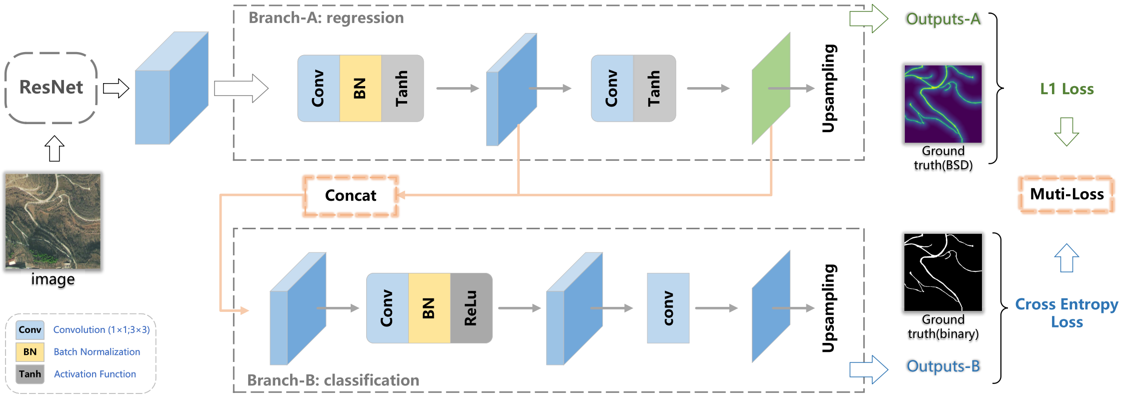

Fig. 1. Overview of the proposed method for road extraction on remote sensing images.

extraction model that combines U-Net with dense connections to alleviate the problem of the tree and shadow occlusion. Zhou *et al.* [13] proposed using dilated convolutions to expand the receptive field of convolution operators and enable richer spatial information. After that, to improve the computational efficiency, Li *et al.* [14] proposed an improved D-LinkNet with a modified D-block, which adds a 1 × 1 convolution layer before the D-block to reduce parameters.

All the abovementioned methods treat road extraction as a binary classification problem where a deep classification model is trained to determine whether a pixel belongs to a road category or nonroad category. However, in WUI areas, the road tends to be narrow and can be occluded by trees or shade. Directly adopting classification-based models can lead to poor performance due to spatial discontinuity. This problem mainly comes from the 0–1 hard classification of each pixel while ignoring the fact that a road network tends to be spatially continuous. To address this problem, we propose to transform the output representation of a binary classification map into a real-value distance map where, for each pixel, our model predicts the distance from this pixel to the nearest boundary pixel. With these real-value distance labels, our model is enforced to learn distance-aware features to help identify road pixels to relieve the spatial discontinuous road problem caused by occlusion, shadow, or roads that are too narrow.

Furthermore, we design a multitask learning architecture to combine the binary classification task with the distance regression task. Especially, a dual-branch network is proposed to predict the semantic classification map and the continuous distance map simultaneously. The semantic classification branch performs conventional binary classification, whereas the distance regression branch generates real-value distance maps. These two tasks are trained simultaneously in an end-to-end manner and produce spatially continuous and semantically accurate road maps.

II. Methodology

A. Boundary Signed Distance

As shown in Fig. 1, two representations of samples are required to be input in our model. This section will first introduce the generation of the boundary signed distance (BSD) masks from the binary masks, and then describe the network architecture. Our BSD representation is inspired by the signed distance transform (SDT) [15]. Given a pixel location x_i, the signed distance value SDT_i is calculated as

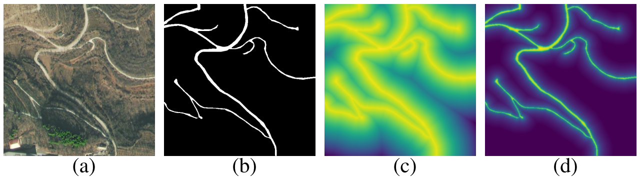
(a) (b) (c) (d)

Fig. 2. Illustration of different output representations. (a) Input image. (b) Binary mask. (c) SDT regression mask. (d) BSD regression mask.

$$\mathrm{SDT}_i = \begin{cases} +\mathrm{Min}_{x_j \in B}\, \mathrm{ED}(x_i, x_j) & x_i \in F \\ -\mathrm{Min}_{x_j \in F}\, \mathrm{ED}(x_i, x_j) & x_i \in B \end{cases} \tag{1}$$

where ED denotes the Euclidean distance and F and B denote the foreground areas (road pixels) and background areas (nonroad pixels), respectively. By this formulation, the signed distance value represents the distance from one pixel to its closest pixels of a different class. The sign here indicates whether the pixels belong to the foreground or background areas.

In this letter, we modified the signed distance to be more suitable for our road extraction model and named it the BSD, which is defined in the following equation:

$$D_i = \underset{x_j \in \mathrm{bound}}{\mathrm{Min}} (\mathrm{ED}(x_i, x_j)) \tag{2}$$

$$\mathrm{BSD}_i = \begin{cases} +H\tanh(\alpha * \log(|D_i|)) & x_i \in F \\ -H\tanh(\alpha * \log(|D_i|)) & x_i \in B \end{cases} \tag{3}$$

where “bound” denotes boundary pixels of the road networks, and α is an hyperparameter. In our formulation, for each pixel location x_i, the distance value (D_i) is defined as the distance from x_i to the closest point x_j located on the road boundary. We feed the distance (D_i) to a logarithmic function to normalize the distance value. In this way, smaller distance values obtain more precise representations, while larger distance values obtain a coarse representation, i.e., more attention is on the road boundary and the surrounding pixels. After that, a HardTanh function (Htanh) along with a scale coefficient α is employed to normalize the distance values to [−1, 1]. By normalization, we can accelerate the model training speed.

Moreover, the BSD of road and nonroad pixels correspond to positive and negative values, respectively, and the value of road boundary pixels corresponds to zero. This will effectively help the model to better distinguish between roads and nonroads.

Fig. 2 shows different output representations, including the binary map, the SDT distance map, and our BSD distance map. From Fig. 2, we can see that unlike binary maps that only indicate road-background category information, the SDT and BSD distance representations can better characterize fine-grained boundary information. These real-value distance representations encourage a model to better recognize object boundaries and provide a more flexible method for identifying road pixels by the to-boundary distance features, rather than only considering the spectral features. In this way, our model tends to correct these ill-shaped roads by comparing the to-boundary distance of these ambiguous pixels with the nearby certain road pixels.

In this letter, we use both a binary mask and a BSD distance map as the output representation of our model. In this way, our model benefits from both 0–1 classification and distance regression tasks. The multitask learning network and optimization process were introduced in Section II-B.

B. Network Architecture

An overview of our proposed method is illustrated in Fig. 1. Our proposed method contains three main components: the backbone network, the regression-classification integration module, and the multitask loss function.

1) Backbone: Our model is built on the ResNet backbone. Considering that WUI roads usually have a small width with only a few pixels, our model is required to maintain the high-frequency details of input images. To achieve this, we change the stride of the last two residual blocks to 1 so that the backbone network can obtain feature maps of the 1/8 size of the input images. Moreover, dilated convolutions are employed to enlarge the receptive field of convolutional layers to enable a larger spatial context. The output feature maps of the backbone network are fed to the following regression-classification integration module for distance-aware feature learning.

2) Regression-Classification Integration Module: As illustrated in Fig. 1, branch-A is designed for the distance regression task, while branch-B is designed for the binary classification task.

In branch-A, a 3 × 3 convolutional layer is first adopted to reduce the dimension of the output feature maps generated by the backbone network. Then, a batch normalization layer is adopted to normalize the distribution of inputs on each minibatch to a standard normal distribution, which can reduce the internal covariate offset and accelerate the learning process [16]. The generated low-level feature maps are designed to characterize both the to-boundary distance features and the semantic features, which are shared across two branches. The low-level feature maps are then fed into a sequence of 1 × 1 convolution layers and activation function Tanh to obtain the low-resolution distance maps. The Tanh function here is used to limit the distance value to the range $[-1, 1]$, which is consistent with the BSD label masks. Note that the L1 loss, as one part of the final loss, will be calculated between the outputs of branch-A and ground-truth BSD masks.

For the classification branch, a direct implementation is to directly feed the low-level feature maps into another decoder branch to produce the classification maps. To enable information flow between these two branches, the low-level feature maps are combined with the low-resolution distance map from branch-A through a concatenate operation. The combined feature maps are then fed into a 3 × 3 convolution layer, a batch normalization layer, and a ReLU activation layer to generate the semantic feature maps. After that, another 1 × 1 convolution layer is employed to obtain the classification score maps. From an overall perspective, the distance regression task in branch-A can be regarded as intermediate supervision for the classification task in branch-B. The score maps are then upsampled to the original resolution to obtain the final classification outputs. The cross-entropy loss, as the other part of the final loss, is calculated between the outputs of branch-B and the ground-truth binary label masks.

C. Loss Function

As mentioned in Section I a multitask training strategy is adopted in our model. The L1 loss is used for the regression task, while the cross-entropy loss is used for the traditional classification task. The loss terms are defined as follows:

$$\mathcal{L}_{\text{bsd}} = -\frac{1}{N}\sum_{i=1}^{N}|\text{BSD}_i - \text{BSD}_i^{gt}| \tag{4}$$

$$\mathcal{L}_{\text{cls}} = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(p_i) + (1-y_i)\log(1-p_i)) \tag{5}$$

where N denotes the total number of pixels, and BSD_i^{gt} and BSD_i denote the ground truth and predicted BSD of pixel i, respectively. y_i and p_i denote the ground truth and predicted label pixel i.

Given these two loss terms, the final loss function of our model is calculated as

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{bsd}} + \lambda * \mathcal{L}_{\text{cls}} \tag{6}$$

where λ is a hyperparameter used to balance the magnitude difference between two types of loss and can also be used to control the proportion of classification and regression tasks.

III. Experiments

To verify the effectiveness of the proposed method for road extraction in WUI areas, we conduct experiments on two data sets and compare the performance of our model with state-of-the-art methods.

A. Data Set

1) Massachusetts: The Massachusetts Roads data set [5] covers a wide variety of urban, suburban, and rural regions, which consists of 1171 aerial images with a resolution of 1.2 m and three spectral bands of RGB, including 1108 images for training, 14 images for validation, and 49 images for testing,

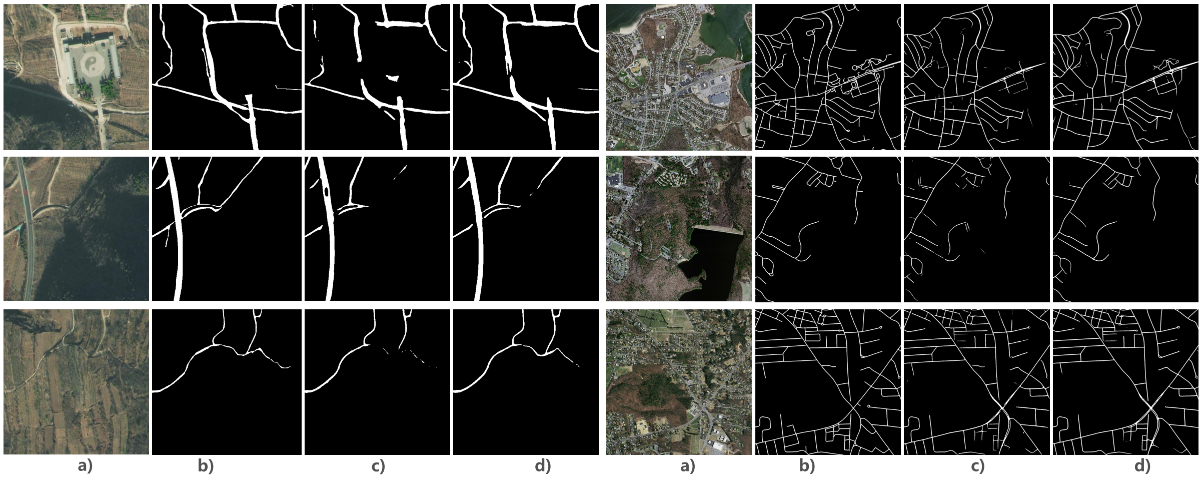

Fig. 3. Selected examples of the road extraction results on the test set of the Massachusetts (left) and WUI-Yajishan data sets (right). (a) Input image. (b) Ground truth. (c) Baseline. (d) Ours.

and each image is 1500 × 1.500 pixels in size. Note that the target labels are generated by rasterizing road centerlines obtained from OpenStreetMap with a line thickness of seven pixels, which has ambiguity in the road boundary area.

2) WUI-Yajishan: To better demonstrate the performance of our model in WUI areas, we self-collect a new data set in Yajishan Mountain, Beijing, and the nearby villages. Unlike the Massachusetts road data set, the roads in our data set mainly cover mountain roads and rural roads, which are generally narrower than urban roads. In addition, in WUI areas, the background is generally more complex, and roads are easily covered by trees or shadows, which makes the road extraction problem more challenging.

Our data set is collected from SuperView-1 images and covers an area of 33 km^2, with a spatial resolution of 0.5 m. The original SuperView-1 images contain four bands of red, green, blue, and near-inferred. In our experiments, we only use three bands (red, green, and blue). We manually vectorized the road objects and then converted them into binary masks. A sliding window strategy is used to generate each image of 512 × 512 pixels in size. Finally, we obtained 1183 images for training with an overlap of 256 pixels and 88 images for testing with no overlap.

B. Implementation Details

We implement our model on four NVIDIA Titan XP GPUs with 12-GB memory and employ a learning rate policy where the learning rate decays according to the cosine function [17] from the initial value to 0.00002. In addition, the initial learning rate and batch size are set to 0.01 and 4, respectively. We train our model for 150 epochs until convergence. For data augmentation, we adopt random left–right flipping for both data sets and random cropping (768 × 768) for the Massachusetts road data set. In addition, α in loss (6) is set to 2. As for the computation, taking the Massachusetts road data set as an example, the total training time is approximately 7.5 h, and the total testing time is approximately 231 s.

C. Evaluation Metrics

The F1 score and intersection-over-union (IoU) are used to evaluate the experimental results. The F1 score is a metric often used to evaluate the binary classification models, which is the harmonic mean of precision and recall. The IoU metric represents the intersection of the prediction and ground-truth regions over their union, and it can be seen as another way to reasonably take both accuracy and recall into account.

In addition, the break-even point with relaxed precision and recall proposed by Mnih and Hinton [5] is also adopted in this letter. The slack parameter ρ is set to 3, which is consistent with previous studies [5], [11].

D. Results

We first conduct experiments to show the effect of using BSD representation on an intermediate regression task and explore the performance of our model with different network depths. To achieve this, we conduct experiments by directly feeding the output of the modified ResNet backbone to the classification task and use it as the baseline model. In this baseline model, only the binary label masks are used as the output representation.

Quantitative results are listed in Table I. From Table I, one can see that our model achieves better performance than the baseline on both data sets. Especially, by using the proposed BSD representation, our model achieves improvements of 4.03% and 3.10% on the IoU for the Massachusetts road data set and the WUI-Yajishan data set. Moreover, using a deeper backbone (Resnet101) further improves the by 1.28% and 0.75% on the Massachusetts road data set and WUI-Yajishan road data set, respectively. We rerun our experiments on both data sets for several times, and the standard deviation is less than 0.5%.

Fig. 3 shows randomly selected examples of the road extraction results on the test set of the Massachusetts and WUI-Yajishan data sets. As shown in Fig. 3, compared with the baseline model, our model produces better road extraction results in both urban areas (Massachusetts road data set) and

TABLE I

ABLATION ANALYSIS ON DIFFERENT DATA SETS. "MASS" INDICATES THE MASSACHUSETTS ROAD DATA SET. "YJS" INDICATES THE WUI-YAJISHAN DATA SET

Data	Method	backbone	precision	recall	F1	IoU
Mass	baseline	ResNet50	82.36	69.97	75.66	60.85
Mass	ours	ResNet50	83.17	72.61	77.53	63.30
Mass	ours	ResNet101	84.65	72.54	78.13	64.11
YJS	baseline	ResNet50	78.64	76.64	77.63	63.43
YJS	ours	ResNet50	78.79	79.42	79.10	65.43
YJS	ours	ResNet101	80.05	78.88	79.46	65.92

TABLE II

COMPARISON WITH THE STATE-OF-THE-ART MODELS. BOLDFACE INDICATES THE MODEL WITH THE BEST PERFORMANCE

Data	Method	Precision	Recall	F1	IoU
Mass	DeepLabV3+	79.16	60.22	67.64	51.95
Mass	D-LinNet+	79.45	71.96	75.15	60.71
Mass	U-Net	84.04	68.90	75.24	60.94
Mass	CDG	81.41	71.80	76.10	61.90
Mass	Ours	**84.65**	**72.54**	**78.13**	**64.11**
YJS	DeepLabV3+	66.80	86.20	75.30	60.40
YJS	U-Net	79.10	75.70	77.30	63.10
YJS	Ours	**80.05**	**78.88**	**79.46**	**65.92**

TABLE III

PARAMETERS COMPARISON ON THE MASSCHUSSETTS ROAD DATA SET

Method	break-even point	parameters
U-Net	0.9053	30.6M
ResUnet	0.9187	**7.8M**
Ours-Res50	**0.9327**	37.4M

WUI areas (WUI-Yajishan data set). More importantly, it can be clearly seen that, by introducing BSD representation, our model generates road predictions with better spatial continuity (e.g., fewer holes and smaller broken length), especially in WUI areas of the WUI-Yajishan data set. This demonstrates that, by incorporating BSD representation as intermediate supervision, our model can successfully alleviate the problem of discontinuous road prediction in WUI areas.

Second, we compare the performance of our model with the state-of-the-art models on the Massachusetts road and WUI-Yajishan data sets. Comparing methods include DeepLabV3+ [10], D-LinNet+ [13], U-Net [8], and coord-dense-global model (CDG) [18]. As shown in Table II, our model achieves the best performance, which verifies the superiority of our proposed model.

Moreover, we show the break-even value and model size of our method, the U-net method, and the ResUnet [11] method in Table III. Results show that our model can obtain better performance than the comparing methods but with more parameters than ResUnet.

IV. CONCLUSION

In this letter, we introduce a method for road extraction from remote sensing images in WUI areas. Unlike existing methods that mostly treat road extraction as a binary classification problem and design deep neural networks for semantic labeling, the proposed method introduces a BSD representation as the fine-grained label representation. In this way, our model is forced to predict the continuous distance representations and, thus, improve the spatial continuities of inferred roads. A multitask network is then designed with the BSD representation as intermediate supervision. Experiments on the Massachusetts road data set and our WUI-Yajishan data set demonstrate the superiority of the proposed method for road extraction in WUI areas. Our proposed method achieves IoUs of 64.11% and 65.92% for the Massachusetts and WUI-Yajishan data sets, respectively, without any postprocessing. Ablation experiments also show that using the proposed BSD representation can effectively alleviate the problem of discontinuous road prediction.

REFERENCES

[1] R. Marikhu, M. N. Dailey, S. S. Makhanov, and K. Honda, "A family of quadratic snakes for road extraction," in *Proc. Asian Conf. Comput. Vis.*, 2007, pp. 85–94.

[2] J. Hu, A. Razdan, J. C. Femiani, M. Cui, and P. Wonka, "Road network extraction and intersection detection from aerial images by tracking road footprints," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 12, pp. 4144–4157, Dec. 2007.

[3] M. Song and D. Civco, "Road extraction using SVM and image segmentation," *Photogramm. Eng. Remote Sens.*, vol. 70, no. 12, pp. 1365–1371, Dec. 2004.

[4] P. P. Singh and R. D. Garg, "Automatic road extraction from high resolution satellite image using adaptive global thresholding and morphological operations," *J. Indian Soc. Remote Sens.*, vol. 41, no. 3, pp. 631–640, 2013.

[5] V. Mnih, "Machine learning for aerial image labeling," Ph.D. dissertation, Univ. Toronto, Toronto, ON, Canada, 2013.

[6] J. Wang, J. Song, M. Chen, and Z. Yang, "Road network extraction: A neural-dynamic framework based on deep learning and a finite state machine," *Int. J. Remote Sens.*, vol. 36, no. 12, pp. 3144–3169, Jun. 2015.

[7] M. Rezaee and Y. Zhang, "Road detection using deep neural network in high spatial resolution images," in *Proc. Joint Urban Remote Sens. Event (JURSE)*, Mar. 2017, pp. 1–4.

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

[9] 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.

[10] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 833–851.

[11] 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.

[12] J. Xin, X. Zhang, Z. Zhang, and W. Fang, "Road extraction of high-resolution remote sensing images derived from DenseUNet," *Remote Sens.*, vol. 11, no. 21, p. 2499, Oct. 2019.

[13] 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. 1–5.

[14] Y. Li, B. Peng, L. He, K. Fan, Z. Li, and L. Tong, "Road extraction from unmanned aerial vehicle remote sensing images based on improved neural networks," *Sensors*, vol. 19, no. 19, p. 4115, Sep. 2019.

[15] Q.-Z. Ye, "The signed Euclidean distance transform and its applications," in *Proc. 9th Int. Conf. Pattern Recognit.*, 1988, pp. 495–499.

[16] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167

[17] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," 2016, *arXiv:1608.03983*. [Online]. Available: https://arxiv.org/abs/1608.03983

[18] S. Wang, H. Yang, Q. Wu, Z. Zheng, Y. Wu, and J. Li, "An improved method for road extraction from high-resolution remote-sensing images that enhances boundary information," *Sensors*, vol. 20, no. 7, p. 2064, Apr. 2020.