Deep Learning-Based Sidewalk Extraction on Aerial Image

Zhibin Bao, Haoxiang Lang, Xianke Lin*
Department of Automotive and Mechatronics Engineering, Ontario Tech University, Oshawa, Canada
*xiankelin@ieee.org

Abstract—High-Definition (HD) maps have drawn lots of attention in the field of autonomous driving in recent years, due to their high localization precision and detailed environmental information. Road network extraction on aerial images, including roads, road boundaries, and sidewalks, is a key component in creating an HD map, which supports navigation and motion or mission planning modules for an autonomous vehicle. In this paper, we present a semantic segmentation-based method to precisely extract sidewalks from aerial images, relying on the strong use of data augmentation. Additionally, we applied batch normalization to the existing semantic segmentation model to increase the training efficiency.

Keywords—semantic segmentation; sidewalk extraction; HD maps; deep learning

I. INTRODUCTION

Autonomous driving has drawn lots of attention in the past few years, especially with the fast development of deep neural networks (DNNs) and electric vehicles. Among all modules in autonomous driving, HD maps have been widely used to provide precise localization ability and environmental information for automated vehicles or robots. As the complexity of modern traffic increases, a precise road network extraction on the aerial image is crucial for HD maps since the automated vehicle localizes itself based on its location on the road and makes path or motion planning based on the road shapes. Recent research on deep learning-based road or road boundary extraction methods has been proven effective in automatically generating road networks from aerial images. Besides roads or road boundaries, the sidewalk is also an important feature on an HD map, especially for small-scale robots or vehicles operating on sidewalks, such as delivery robots. Since many sidewalks are beside the motorway, it is critical to precisely present the sidewalks or sidewalk boundaries on the HD map to prevent robots from deviating from sidewalks or autonomous vehicles from driving onto sidewalks. Most recent publications and research on sidewalk extraction mainly focus on sidewalk semantic segmentation on street view images [1]–[4]. Current state-of-the-art methods of feature extraction from aerial images also mainly focus on road/road boundaries [5]–[10], road furniture [11], and road markings [12]–[14]. However, the current research still lacks efficient methods for extracting sidewalks from aerial images. In this paper, to fill the gap in the current sidewalk extraction research, we proposed a sidewalk extraction from aerial images method, based on U-Net [15], to automatically extract sidewalks using semantic segmentation. The novelty and contributions of the proposed work are summarized below:

1. We replaced the traditional manually labeling the sidewalk network method by applying deep learning method for sidewalk extraction.
2. Data augmentation was applied to the training dataset to increase the size of the dataset and avoid model overfitting.
3. Batch normalization [16] was added to the original U-Net model to increase the training efficiency.
4. This work aims to break the gap in the research on sidewalk extraction on aerial images and provide the community with a jumpstart on this research field.

The rest of this paper is organized as follows. Section II reviews the related work. Section III discusses the methodology of the proposed method. The performance of the proposed method and its limitations are discussed in Section IV. Finally, Section V concludes this paper.

II. RELATED WORK

Sidewalk extraction is categorized under road network extraction on aerial images, which is one of the general and vital steps in creating an HD map. Road network extraction on aerial images was traditionally done by labor, which was costly, time-consuming, and low precision due to human errors. In recent years, deep learning-based road network extraction methods have been developed and widely used to increase HD map precision and reduce the amount of manual work. Current road network extractions are typically done in three methods: segmentation-based methods, iterative graph growing methods, and graph-generation methods.

A. Segmentation-based Method

Segmentation-based methods extract the road network from aerial images by first predicting the probabilistic segmentation map from the aerial image. The prediction is then refined and
used to extract the road network graph through post-processing. Mattys et al. proposed DeepRoadMapper [5] to directly estimate the road topology and extract the road network from aerial images. The author also applied A* algorithm to solve the road discontinuity issue. [7] and [8] also applied segmentation-based methods to extract the road network from aerial images and solve the discontinuity issue. The former predicts the orientation and segmentation of the road network and corrects the segmentation results using convolutional neural networks (CNNs), and the latter adds an edge detection algorithm on top of the segmentation result.

B. Iterative Graph Growing Method

In iterative graph growing methods, several vertices of the road network are selected, and the road is generated vertex by vertex until the entire road network is completed. Bastani et al. utilize an iterative graph construction process to extract the road networks from aerial images [9]. The approach, RoadTracer, starts from a known single vertex on the road network and applies a CNN-based search algorithm to add vertices and edges to the road network iteratively. The search algorithm continuously explores until the road network is completed. This approach has been proven effective in extracting road networks from aerial images and solving road discontinuity issues caused by occlusions (trees, buildings, shadows).

C. Graph-Generation Method

Graph-generation methods utilize an encoder-decoder algorithm to extract the road network from aerial images. It first encodes the input aerial images into vector fields for neural networks to learn road network features for prediction. The decoder algorithm decodes the prediction from the encoder branch into road network graphs. The graph-generation methods have been widely used to predict road network graphs, including line-shaped objects [10], road boundaries [6], line segments [17], and polygon-shaped buildings [11].

The methods mentioned above have made great progress in road network extraction on aerial images. However, their works still mainly focus on extracting roads or road boundaries. When creating an HD map for autonomous driving, the sidewalk is
also a necessary feature to provide the automated vehicle with more information about the environment. Thus, methods for precisely and automatically extracting sidewalks from aerial images are in demand.

III. METHODOLOGY

In our proposed method, we achieved sidewalk extraction from aerial images using the semantic segmentation method. The goal of semantic segmentation is to correctly predict each pixel of an image as a corresponding category of what is being represented. When training a deep neural network (DNN) model for semantic segmentation, the training data usually contains two types of images, the original image and the mask image. In the road network extraction application, the original image refers to the aerial image, and the mask image refers to the road network annotation. A sample of an aerial image and its sidewalk mask is presented in Fig. 1. Image augmentation is applied before training the model to increase the number of aerial images and sidewalk masks, including horizontal flip and vertical flip, as shown in Fig. 2. The Random Rotation feature is also applied on the dataset while preparing the training script in PyTorch.

The DNN model used in our sidewalk extraction method, U-Net [15], was proposed by Ronneberger et al., and was originally designed for biomedical image segmentation. It utilizes an encoder-decoder architecture with the bottleneck layer at the bottom of the “U” shape. The encoder branch down samples the input aerial image into feature maps. At each encoder level, the output feature map is saved and later used in the decoder branch for concatenation. The decoder branch upsamples the feature map to regain the same resolution as the input aerial image. At each decoder level, the output feature map saved from the same encoder level is concatenated with the new feature map obtained in this decoder level before upsampling. At the network’s output level, a one-by-one convolutional layer, followed by the Sigmoid activation function, is used to convert the feature map into a probability distribution. A threshold of 0.5 is used here, indicating that if the output probability is greater than 50%, the prediction will be the sidewalk; otherwise, the prediction will be the background. The probability distribution output will then be saved as an image format using the torchvision library provided by PyTorch. The structure of the U-Net used in this paper is shown in Fig. 3.

The input of the architecture is the satellite image, and the output is the sidewalk predictions. After each training epoch, output segmentation predictions are compared with ground truth labels to compute the training loss. The training loss is then used to update the weight to minimize the loss. In addition to the original U-Net model, we added a batch normalization layer after each convolutional layer (3x3) to increase training efficiency. Batch normalization is a pre-processing technique that normalizes different sources of data within the same range. It is done on mini-batches instead of the entire dataset, allowing the training to have higher learning rates and making the learning easier. The model was trained on the NVIDIA RTX 3080Ti GPU, and the training parameters are listed below:

- Batch size = 1
- Number of epochs = 10
- Learning rate = 0.0001 (1e-4)
- Train size : validation size = 9:1

Accuracy and Dice Score are used to evaluate the model performance. The two evaluation metrics are shown in (1) and (2):

\[
\text{Accuracy} = \frac{\text{num of correctly predicted pixels}}{\text{num of all pixels}} \quad (1)
\]

\[
\text{Dice Score} = \frac{2 \times TP}{2 \times TP + FN + FP} \quad (2)
\]
where \( TP \) is true positive, \( FN \) is false negative, and \( FP \) is false positive.

IV. RESULTS AND LIMITATIONS

After 10 epochs of training, the model reached an accuracy of 99.27\% and a dice score of 0.8064. The sidewalk segmentation results are shown in Fig. 4, along with the original aerial images and ground truth labels. The trained model was able to predict most of the sidewalk pixels correctly from the aerial image. Additionally, we also trained the ResUNet [18] model on the sidewalk dataset using the same training parameter. The evaluation results are shown in Table I. Both models achieved similar accuracy, however, the UNet model has a higher dice score, indicating that UNet has a better performance on the sidewalk dataset.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Dice Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNet</td>
<td>99.27%</td>
<td>0.8064</td>
</tr>
<tr>
<td>ResUNet</td>
<td>99.28%</td>
<td>0.7851</td>
</tr>
</tbody>
</table>

Although the model has a fairly high accuracy of predicting the sidewalk, one issue still limits the performance of the sidewalk extraction, which is the occlusion over sidewalk that is caused by trees, buildings, and shadows. An example of such occlusion is shown in Fig. 5(a). In Fig. 5(a), the occlusion is highlighted in the red box, where a part of the sidewalk is covered by a tree. The ground truth label of this aerial image is shown in Fig. 5(b), where the occlusion part is annotated as the sidewalk label. However, the sidewalk segmentation prediction of this aerial image does not predict the occlusion part as the

![Aerial Image](image1.png)

![Ground Truth](image2.png)

![Prediction](image3.png)

**Fig. 4.** Sidewalk segmentation results. Images on the left column are the original aerial images, ground truth labels are in the middle column, and the segmentation predictions are on the right column.

![Aerial Image](image4.png)

![Ground Truth](image5.png)

![Prediction](image6.png)

**Fig. 5.** Sidewalk segmentation discontinuity issue caused by occlusions. (a) is the aerial image, (b) is the sidewalk ground truth, and (c) is the segmentation prediction.
sidewalk, which can be observed in the red box in Fig. 5(c). When the sidewalk segmentation is used for the HD map, such segmentation imperfection can directly affect the mission/motion plan of the autonomous vehicle/robot. The system can potentially treat such discontinuity as a dead end or an opening on the sidewalk, which can lead to a stop or wait command instead of a move forward command. It is necessary to connect the discontinuity to avoid giving a false command to the autonomous vehicle/robot. The discontinuity issue caused by occlusions can potentially be solved by adding a path-planning algorithm to complete the sidewalk segmentation. Another potential solution for solving the discontinuity issue can be giving higher bias or weights to the sidewalk that has occlusions during model training. This allows the neural network to pay more attention to such occluded sidewalk features. Additionally, Conditional Random Fields (CRF) [19] can also be applied to refine the segmentation results.

V. CONCLUSION

This paper presents a sidewalk extraction method based on U-Net to fulfill the gap in road network extraction research. The proposed method achieved 99.27% accuracy in predicting sidewalk pixels. The limitation of the proposed method is the sidewalk discontinuity issue caused by occlusions on top of the sidewalk. Thus, the future work of this research is to solve the discontinuity issue of sidewalk segmentation.

ACKNOWLEDGEMENT

This study was supported by start-up funding under Dr. Xianke Lin.

REFERENCES


