Road Segmentation for Pedestrian Trajectory Prediction

Ang Xiao

May 31, 2023

1 Statement of the Problem to be Solved

The central problem of this project is to develop an effective machine learning model capable of identifying walkable road areas from aerial or drone-captured images. This function is instrumental in predicting pedestrian trajectories, which can greatly enhance navigational safety and efficiency.

2 Review of the State of the Art

Traditional methods primarily involve the manual identification of walkable areas, a process which can be time-consuming and subject to human error. Recent advancements in machine learning have enabled automated identification. A prominent example of such advancements is the Stanford Drone Dataset [1]. However, this dataset is originally composed of videos, requiring conversion into a usable format for our specific problem.

3 Proposed Solution

Our proposed solution is to develop a machine learning model using U-Net architecture [2], a design particularly suited for segmentation tasks. It is built with simple, repetitive blocks - including convolutions, ReLu, and max pooling for the downsampling/encoding path, along with upsampling and ReLu for the upsampling/decoding path - which makes it relatively straightforward to implement. One of U-Net's most significant advantages is its effectiveness with a small amount of training data. Despite the modest size of the dataset we used, U-Net's design facilitates good performance. It has proven its capabilities in various benchmarks and continues to rank decently in Kaggle segmentation challenges. The model will be trained on drone-captured images to identify walkable road areas, and mask images [3] will be incorporated to guide the learning process.

4 Methodology

The Stanford Drone Dataset [1] was utilized for training and testing our model. The original videos were converted into still images, resized to 512x512 pixels, and padded with zeros to ensure ease of processing. The dataset was partitioned into two subsets: 43 images for training and 17 images for testing.

The model employs a U-Net based architecture, renowned for its efficacy in image segmentation tasks. It incorporates multiple downsampling and upsampling stages, along with convolutional layers at each stage. A dropout layer is included in the final downsampling stage to prevent overfitting.

5 Detailed Description and Analysis

The U-Net architecture is not only suitable for our application but also well-equipped to handle the small size of our training dataset. Its simple structure is constituted of repeating building blocks, including convolutions, ReLu activation, and max pooling for the downsampling path. The upsampling path similarly comprises of upsampling, convolutions, and ReLu activation.

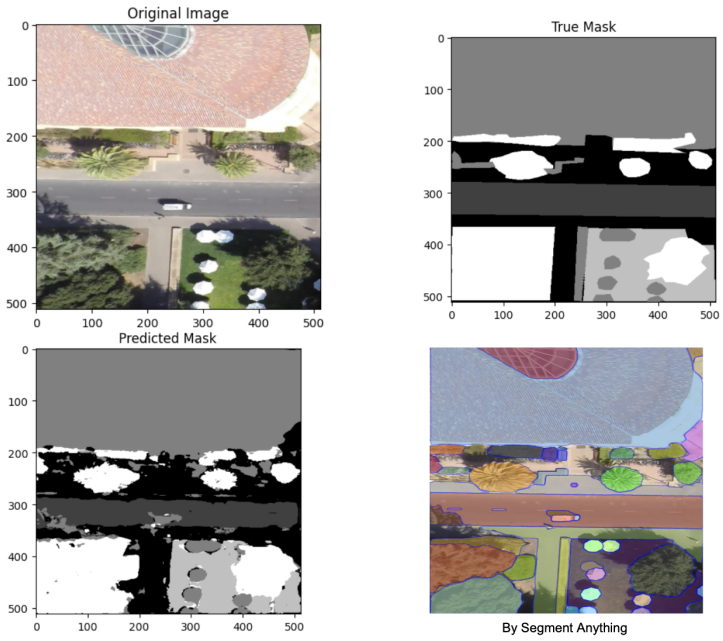

Figure 1: Example of prediction mask for walkable road areas.

Upon examining Figure 1, we note that the prediction mask is capable of making reasonable predictions about where people can walk. However, the

model falls short when attempting to discern detailed textures, and the simplification of the image by the True Mask leads to some loss of information.

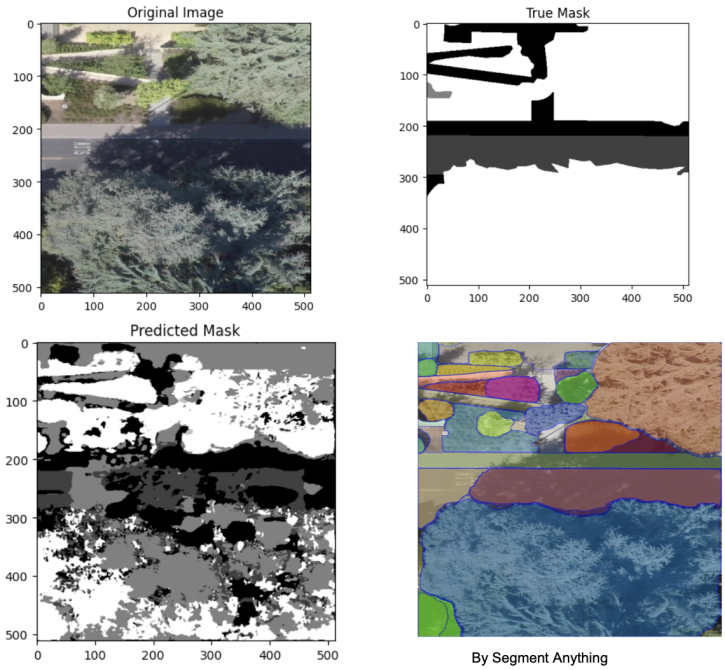

Figure 2: Example of complex details prediction challenges.

In Figure 2, it is observed that complex details such as tree textures are challenging for our model to predict accurately. The model lacks ground truth understanding, such as the fact that people can probably walk under green trees but not on green grass, and people are unlikely to walk on roofs. This is a limitation of supervised learning, which we attempted to address using mask information with limited success.

6 Results

The dataset used for this project, containing only 60 images, was relatively small. As a result, we limited the training of the model to 100 epochs to prevent overfitting, as shown in Figure 3 with the loss function plot. Despite this limitation, the model demonstrated a capacity to predict walkable areas, albeit with some challenges:

- Difficulty in discerning detailed textures such as trees and shadows.
- Challenges in distinguishing between roofs and the ground in aerial images.

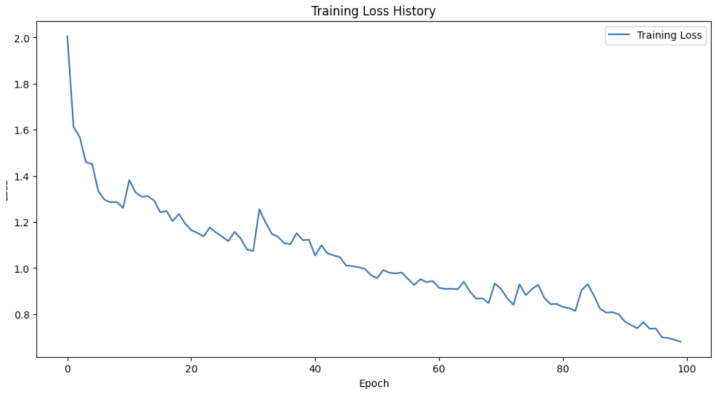

Figure 3: Training Loss History

- Difficulty in identifying obscured walkable areas under trees or buildings.

These results suggest that our U-Net-based model can be effectively applied to recognize walkable road areas in drone-captured images, provided these identified challenges are addressed in future iterations.

7 Conclusions

Despite the challenges, our U-Net based model demonstrates promise for the automated recognition of walkable road areas in drone-captured images. For future work, it is suggested to apply transfer learning techniques to enhance model performance. Pre-training the model on a larger dataset, like ImageNet, and then fine-tuning it on our specific task using our smaller dataset could potentially yield better results.

References

[1] A. Robicquet, A. Sadeghian, A. Alahi, S. Savarese, "Learning Social Etiquette: Human Trajectory Prediction In Crowded Scenes," in European Conference on Computer Vision (ECCV), 2016.

[2] O. Ronneberger, P. Fischer, T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015.

[3] K. Mangalam, Y. An, H. Girase, J. Malik, "From Goals, Waypoints & Paths To Long Term Human Trajectory Forecasting," arXiv preprint arXiv:2012.01526, 2020.

[4] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao et al., "Segment anything," arXiv preprint arXiv:2304.02643, 2023.