Semantic Segmentation on Aerial Drone Images using U-Net

Star Chen, Muyang Xu, Chengyu Zhang

**Introduction:**

Aerial Drone, or unmanned aerial vehicle (UAV), which came in as crucial military assets for scouting, has recently gained an increased popularity in civil use. One key feature of aerial drone is its ability to produce high quality nadir view images and videos through the camera attached to the drone. For popular manufacturers such as DJI, while there are already certain automations in drone piloting and image capturing, they are still relatively basic and human interference is required in the two fields.

To further simplify the drone operation and improve customers experience, we deem that further automation is needed. To start with, object detection in aerial drone images is vital for both drone piloting (e.g., identifying and avoiding obstacles during flight) and image capturing (e.g., identifying key objects and locking camera views during flight).

That’s when machine learning strategies should come in, specifically, image semantic segmentation, where given images, the algorithm is able to identify different objects in those images and apply colored masks to objects. In this way, the previously proposed improvements in drone automation could be achieved. While image semantic segmentation is relatively ripe, most are implemented for daily scenes shot on usual view directions and few are on the top-down birds eye view images. Therefore, in this project, image semantic segmentation, specifically U-Net, would be applied to the aerial drone images to get an ideal object detection in the nadir view images.

**Objectives:**

For the dataset, this project will use the “[Semantic Drone Dataset](https://www.tugraz.at/index.php?id=22387)” found on the Institute of Computer Graphics and Vision. The dataset consists of 400 images of different houses from nadir (bird's eye) view acquired at an altitude of 5 to 30 meters above ground. It is created initially for the understanding of urban scenes for increasing the safety of autonomous drone flight and landing procedures. Since our project intended to do object detection for drone images in civil use, which is mainly in neighborhoods, this dataset is ideal to achieve such a task. The dataset is detailed into three parts: original full color images of high resolution (6000\*4000 pixels), RGB masks for each of the images which different colors of the mask would cover their corresponding different objects in each image, a csv file that gives the information of the labels each of the RGB colors represent (there are 24 labels such as “person”, “door”, “wall”).

For the preprocessing, since in an image semantic segmentation architecture, the mask of an image must contain the code of the class for that object, each pixel in that image should be represented with a class number in the mask, in our case, from 0 to 23. So, our data preprocessing would mainly involve the transformation of the RGB masks into greyscale masks. Since the 400 image datasets is not divided into training and testing set, we would have to manually split the dataset with the initial ratio be 0.8:0.1:0.1 (train:val:test), subject to be tuned. While we are not sure if the 400 images would suffice to generate an ideal result, k-fold cross validation may well be applied.

Our image semantic segmentation approach would mainly follow the U-Net architecture proposed in “[U-Net: Convolutional Networks for Biomedical Image Segmentation](https://arxiv.org/abs/1505.04597)”. The reason to choose U-Net is based on its claimed high performance upon limited datasets. Also, as discussed before, the aerial drone images are shot from a bird’s eye view, different from what’s usually seen. While U-Net is originally applied on the biomedical image shot from microscope, we found that an analogy could be drawn: both image sets are a top-down view of the objects, which might make U-Net more useful in the aerial image segmentation task. Alternatively, if the result is not ideal, we would try to also incorporate the PSPNet in “[Pyramid Scene Parsing Network](https://arxiv.org/abs/1612.01105)”, where the context of the images is emphasized through its pyramid structure.

Moreover, since each of the image is relatively high resolution (6000\*4000 pixels), we might have to tune our batch size for our training epochs. Since the original U-Net already uses single batch size, in order to reduce overhead, we might also try to reduce the resolution of the original images and the masks.

Since the original implementation given in the U-Net paper is based on Caffe, we would try to implement the U-Net structure using PyTorch basic modules (and the whole project would be based on PyTorch framework).

The mainly loss function in our project would also follow the U-Net paper, using pixel-wise soft-max combined with the cross-entropy loss function. Pixcal-accuracy (PA) and Intersection over union (IoU) would also be applied in training and evaluation.

For the evaluation, as discussed, would be based on the previously split test set. Strategies are mentioned above. Moreover, we would like to generate our own testing data, with other high resolution drone footages found on the Internet or shot by ourselves. The tool to manually apply a RGB mask would be the [Semantic Segmentation Editor](https://github.com/Hitachi-Automotive-And-Industry-Lab/semantic-segmentation-editor) on GitHub. Through this way, we hope that we could test the generality of our approach, as well as getting ourselves the experience and knowledge of generating our own dataset.

Through this project we hope to: generate an ideal result (high accuracy and robustness) towards the semantic segmentation in aerial drone images, have an understanding of the concepts and application of image semantic segmentation, as well as the structure of a semantic segmentation algorithm, the experience of generating dataset and lastly, an end-to-end machine learning project experience.

**References:**

Dataset Resource: Semantic Drone Dataset <https://www.tugraz.at/index.php?id=22387>.

Alternative Dataset Resource: Aerial Semantic Segmentation Drone Dataset <https://www.kaggle.com/bulentsiyah/semantic-drone-dataset>.

Semantic Segmentation Paper (U-Net): U-Net: Convolutional Networks for Biomedical Image Segmentation. Ronneberger et al., 18 May 2015, <https://arxiv.org/abs/1505.04597>.

Additional Semantic Segmentation Paper (PSPNet): Pyramid Scene Parsing Network. Zhao et al., 27 Apr 2017, <https://arxiv.org/abs/1612.01105>.

Evaluation Data Generator: Semantic Segmentation Editor <https://github.com/Hitachi-Automotive-And-Industry-Lab/semantic-segmentation-editor>.