Research Summary

Fred Lu

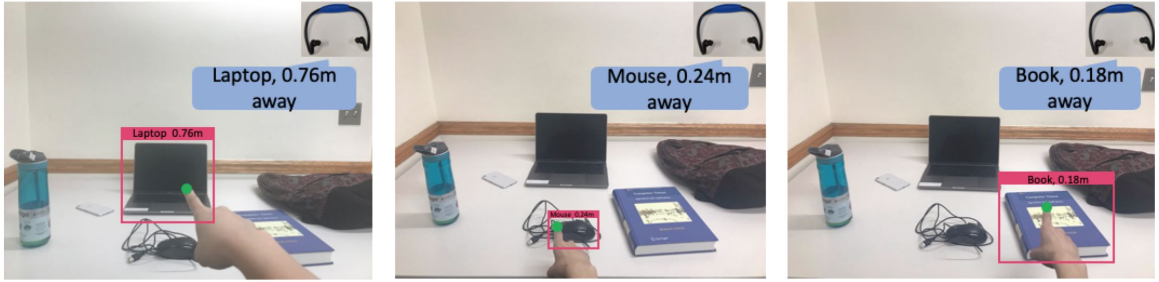
1. Weakly Supervised Point-to-tell

Background

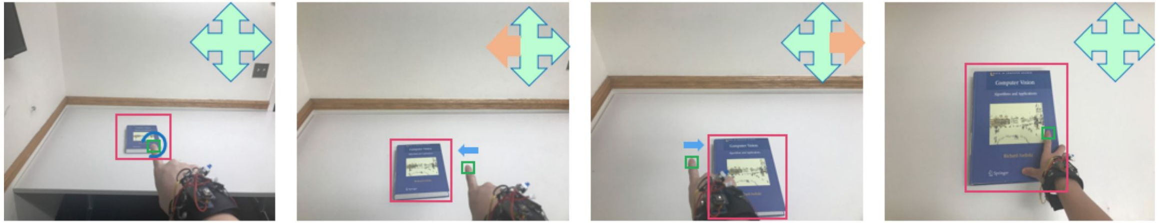
Wenjun Gui and others from the NYU MMVC Lab has investigated an haptic feedback system to assist the visually impaired better navigator around indoor space.

However, the system required training images to be annotated with bounding boxes, which could be costly. My research project aims to find out if weak supervision can yield an effective neural network model that accomplishes point-to-tell.

Point-to-tell Demonstration



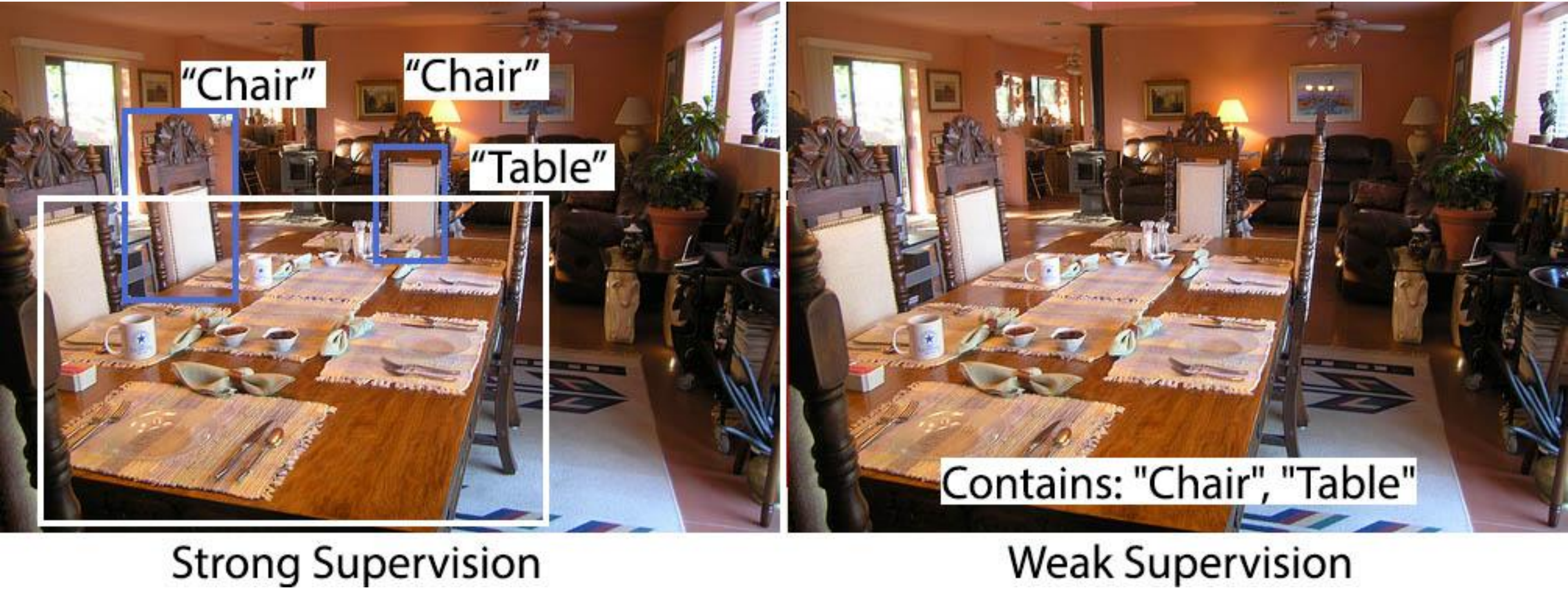
Point-To-Tell



Point-To-Touch

Both point-to-tell and point-to-touch relies on an accurate object detection network to determine the locations of objects and the user’s hand.

Idea: Weak Supervision



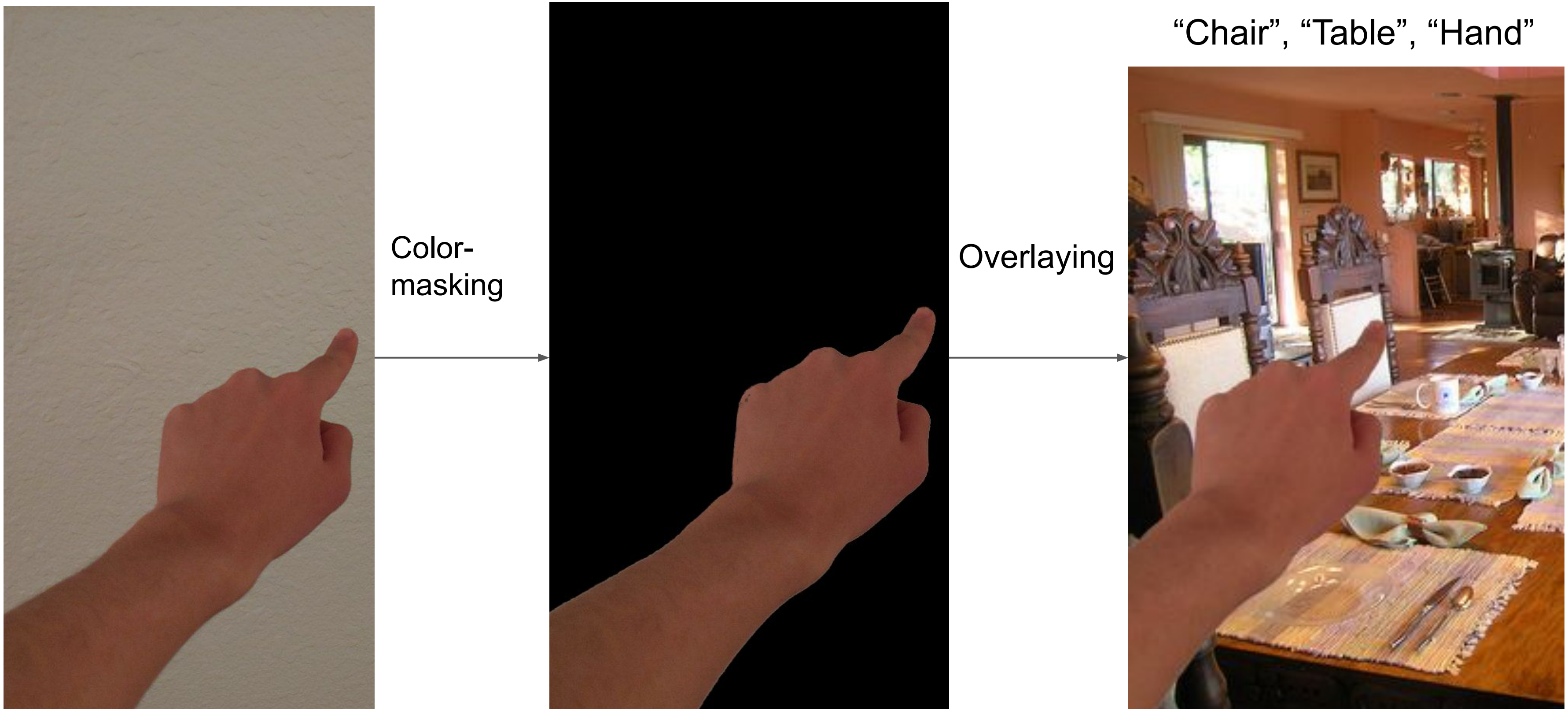
In weakly supervised training, the model only needs to know what object classes are present in an image.

Synthesizing Dataset

We want to base the dataset on the PASCAL VOC SBD dataset as it is widely used and contains several object classes that are useful for in-door navigation.

However, we need to incorporate samples of hands in the dataset as well.

Synthesizing Dataset



Color-masking

Overlaying

“Chair”, “Table”, “Hand”

Synthesizing Dataset

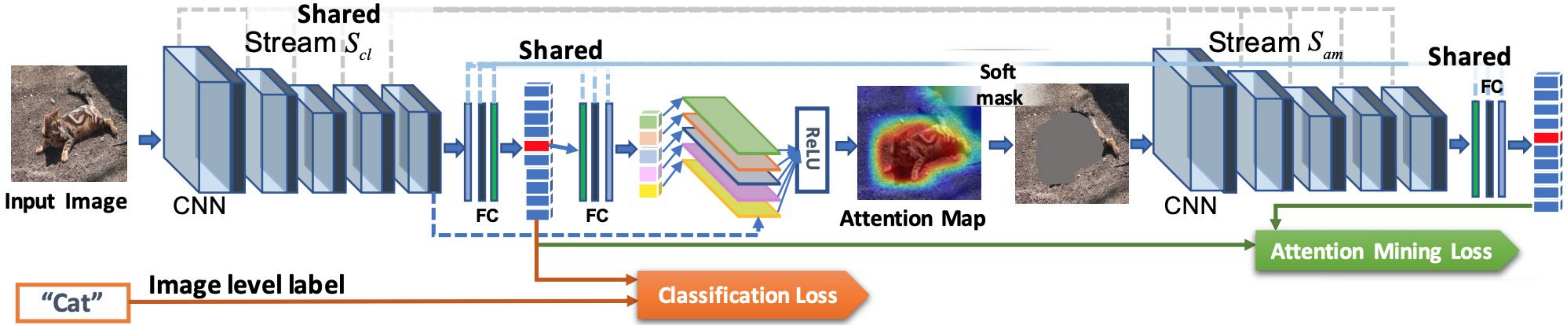
For every image from the VOC SBD, to balance the number of positive and negative hand samples, we overlay a pointing hand at each of the objects we’re interested in and save the new images along with the same number of copies of the original image. (So that a hand appears in exactly 50% of the synthetic dataset).

As such, we ended up with 12656 images for the synthetic dataset.

Model Architecture

The model itself has a convolutional structure with a few fully connected layers attached to output a classification vector for all classes.

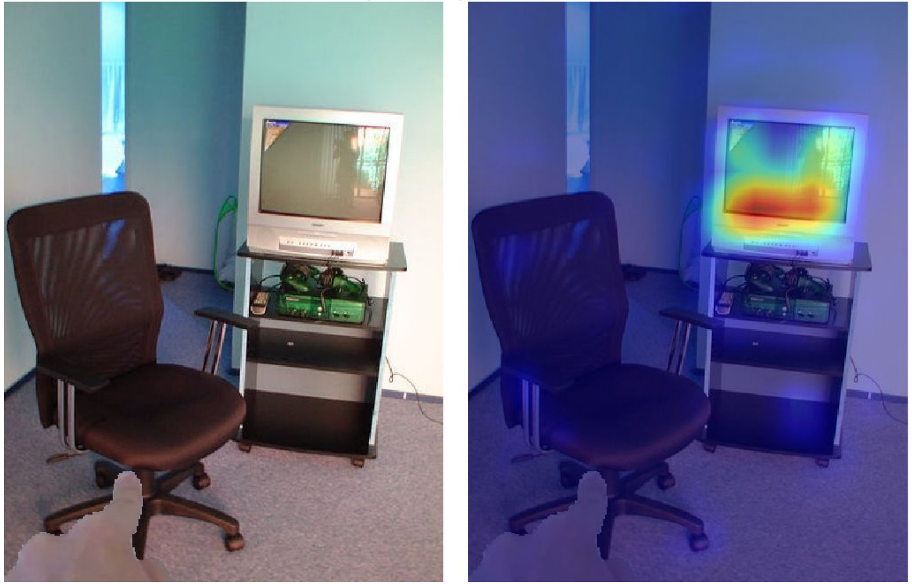
We train the model using the classification loss and the attention mining loss. The latter tells the model to produce attention maps that cover the most distinctive parts of the object.



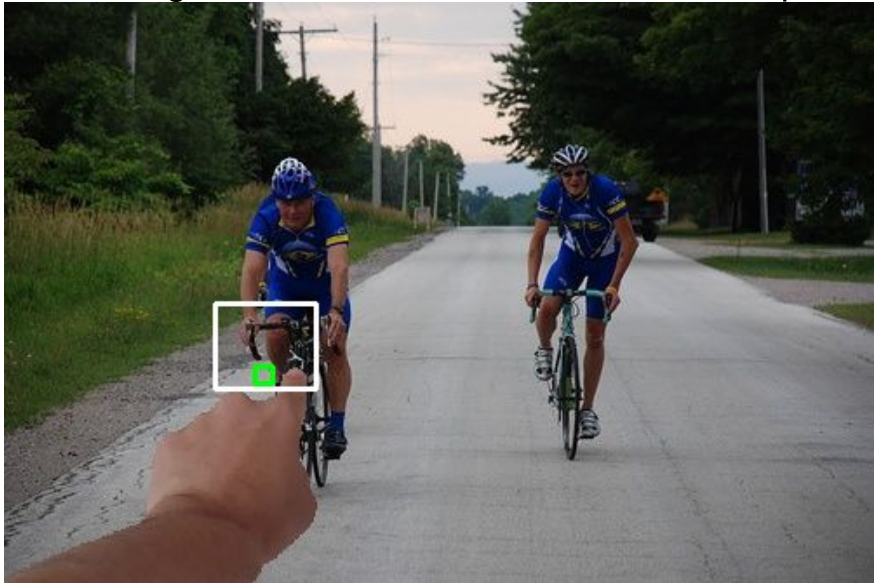
Inference

We use the classification output to determine what objects and if a hand is present in an image. We then use the attention maps to localize all present objects.

Attention Map Output for TV



Bounding Boxes calculated from attention Maps



Conclusion

Using weakly annotated images to train an image classification network results in a usable object localization ability, which can be used to implement both point-to-tell and point-to-touch with a significantly reduced time and money cost of annotating training images.

2. Few-Shot Aerial Photo Segmentation

Background

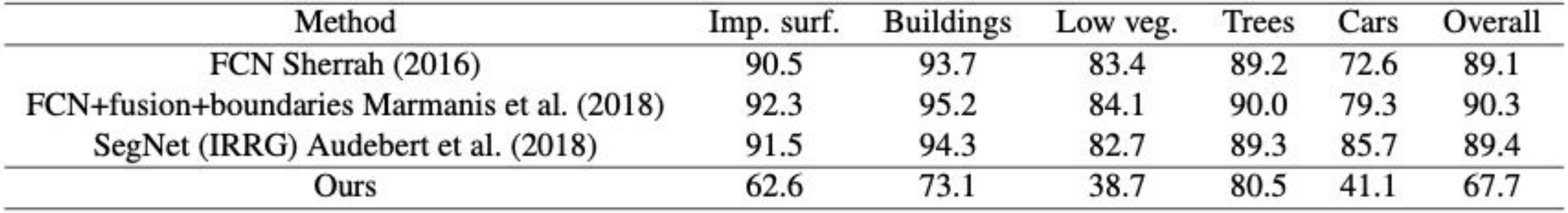
Segmenting aerial or satellite photos using neural networks presents a challenge in obtaining sufficient training data because there are not many public datasets of segmented aerial photos available.

- We want to investigate if using just a few ground-truth patches of aerial photos can produce semantic segmentations of similar quality compared to conventional approaches that require vast amounts of training data.

Background

Xiang Li’s findings have already suggested that using only a few ground-truth aerial images led to acceptable results.

Table - fully trained models versus our approach trained with 5 417x417 patches:



Method	Imp. surf.	Buildings	Low veg.	Trees	Cars	Overall
FCN Sherrah (2016)	90.5	93.7	83.4	89.2	72.6	89.1
FCN+fusion+boundaries Marmanis et al. (2018)	92.3	95.2	84.1	90.0	79.3	90.3
SegNet (IRRG) Audebert et al. (2018)	91.5	94.3	82.7	89.3	85.7	89.4
Ours	62.6	73.1	38.7	80.5	41.1	67.7

Even Weaker Ground Truth

But we can further reduce the amount of annotation required!

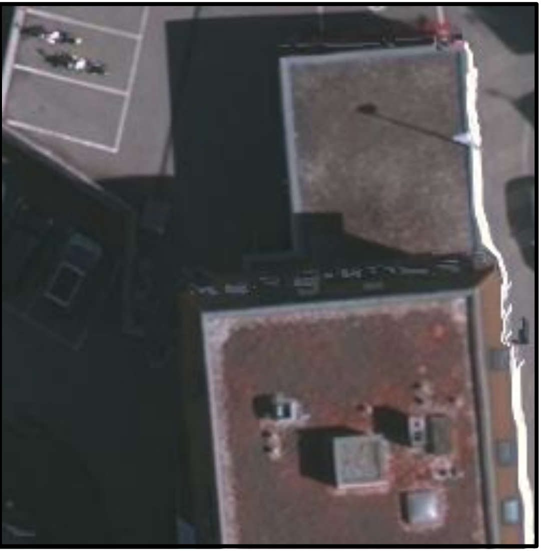
Instead of using pixel-level semantic segmentations, we can try using “bounding boxes” or “scribbles”.

- Either method will allow for an even quicker and easier way to create training data for out model.

Bounding Boxes

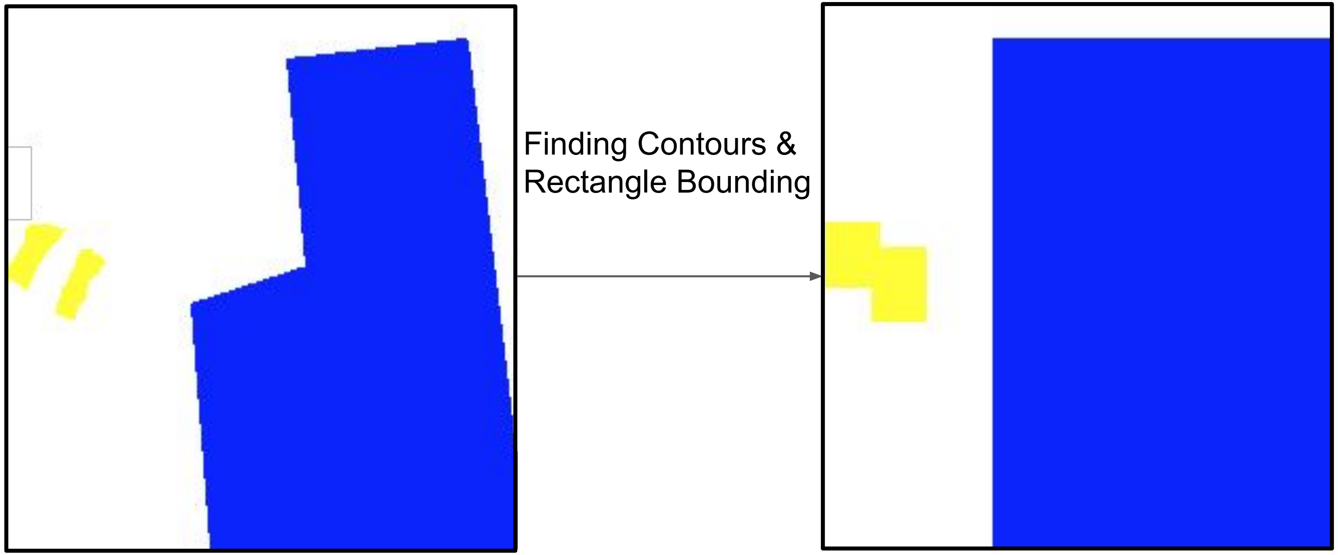
- Easy Annotation but contains unwanted pixels

Aerial Photo



Cars and Building: Regular Segmentation

Cars and Building: “Bounding Box” Segmentation

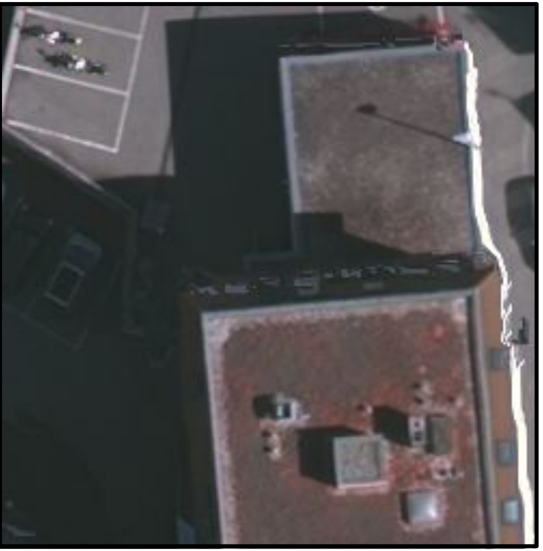


Finding Contours & Rectangle Bounding

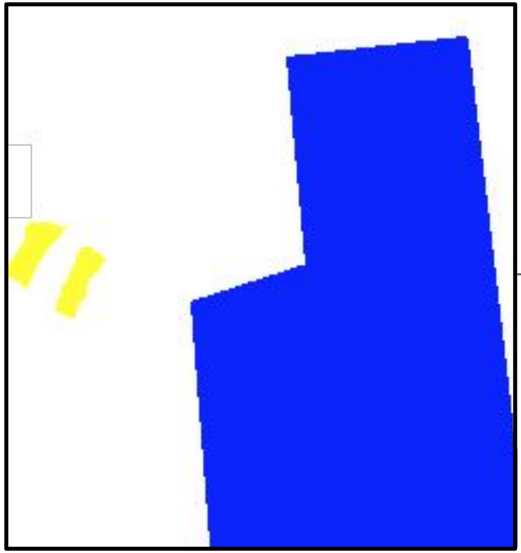
Scribbles

- Easy Annotation but doesn’t contain all pixels belonging to the object

Aerial Photo

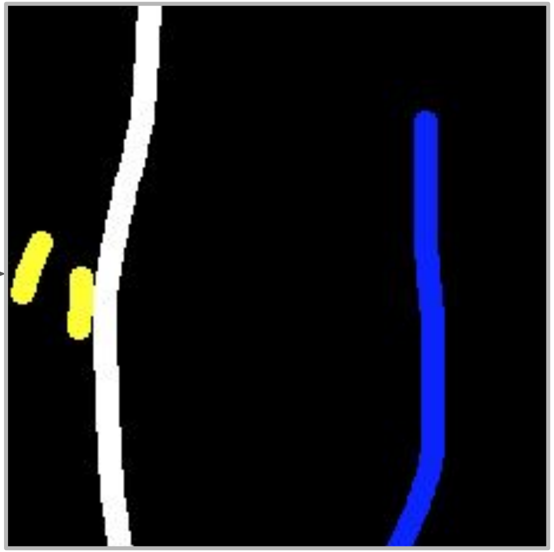


Cars and Building:
Regular Segmentation



Manual Annotation

Cars and Building:
“Scribble” Segmentation

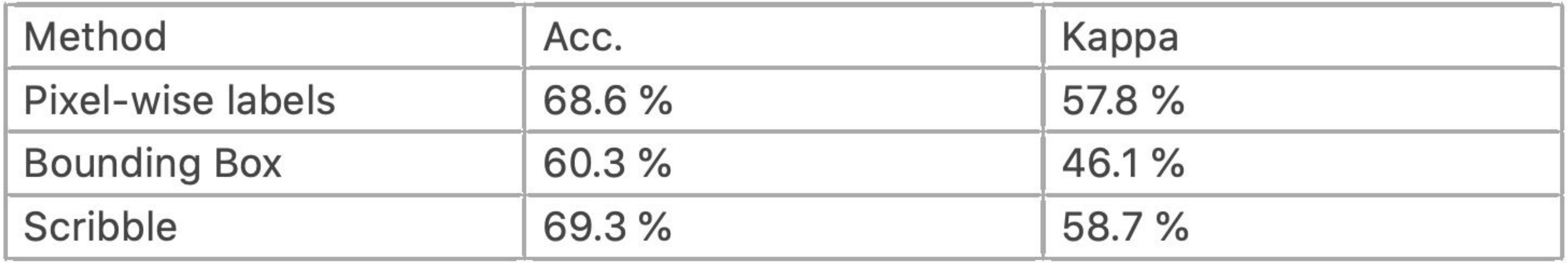


Test Results

By running three experiments -- using regular, “bounding box”, and “scribble” training images -- we found using weaker annotations leads to a similar level of performance of 5-shot segmentation.

“Scribble” achieved better accuracy despite being a weaker annotation than pixel-wise labels. The increase was likely due to the differences in the selected training patches.

“Bounding Box” was the easiest to annotate and achieved slightly worse accuracy.



Method	Acc.	Kappa
Pixel-wise labels	68.6 %	57.8 %
Bounding Box	60.3 %	46.1 %
Scribble	69.3 %	58.7 %

Conclusion

With the PANet (Prototype Alignment Net) model, aerial image segmentation can be accomplished with only a few annotated patches.

Using weaker annotations such as “bounding box” or “scribble” can further simplify the creation of training data, with some or little reduction in segmentation accuracy.

Weaker annotations could be useful when the time or monetary budget of obtaining training data was limited and perfect accuracy was not required.