The dataset is composed of several sequences of videos produced by a camera on a moving platform. Each frame has hand labelled annotations denoting bounding boxes of pedestrians. Overall there are seven sequences of videos with a combined 4,534 frames. The data was split into a training and test set where we have the frames from five sequences (3,105 frames) in the training set and two sequences (1,429) in the test set as shown in Table 1. The annotations are not complete in that they do not strictly label all pedestrians in a given image. It is usually the case that only pedestrians which take up a certain subjective threshold of the screen are labelled. This leads to what could be some false negatives in the training set.

The complete pipeline from video frame to bounding box output is shown in Figure 6. We start with a given video sequence and split it up by frames. Then we run an algorithm to generate proposal bounding boxes, in this case we use selective search, which we cache for use across the process. We pass these bounding boxes along with the original image to the detector which is our convolutional neural network. The CNN produces softmax scores for each bounding box which are used in the final non-maximal suppression step.

We find that Alexnet with pretrained weights from Imagenet performs the best in moderate prediction score thresholds while a slightly larger variant of Alexnet, trained from scratch, performs better at higher acceptance thresholds. There is plenty of future work to be done to improve these results. The main limiting factor was GPU support which could have significantly sped up computation as shown in Figure 7. Additional tuning of hyperparameters such as overlap thresholds may also help.