* **Motivation**: What problem are you tackling?

Most businesses and governments use CCTV footage to prevent and catch crime. However, a core issue with this footage is the vast amount of it that exists. Traditionally, many hours of footage need to be manually combed over to identify when a particular suspect entered the frame of the camera. Our goal is to use just a text description of a criminal to find that criminal in CCTV footage.

* **Method**: What machine learning techniques are you planning to improve upon?

One potential approach to solve this problem would be to break the footage into frames and then, from each frame (with some set interval between frames), use a segmentation model to identify each individual person in the frame (such as YOLO). Then, you could show each person captured throughout the entire footage, which would allow reviewers to comb through the footage much quicker.

We plan to improve beyond person segmentation from previous approaches. In addition to capturing each person, we plan to also match them to a description. In order to do this, we plan on using an aligned text and vision transformer model (such as CLIP) to create embeddings of both the text descriptions and the images of people captured and then use some similarity function (such as cosine similarity) to match the closest images of people to the text description.

* **Intended experiments**: What experiments are you planning to run? How do you plan to evaluate your machine learning algorithm?

In order to evaluate the effectiveness of our approach, we can describe a person in the cctv footage, and then run our vision and text transformer models to rank each person based on how close the model predicts them to be the person described. Then, we can identify the rank the actual person described placed among every person segmented, in order to figure out how good our approach is at identifying criminals.