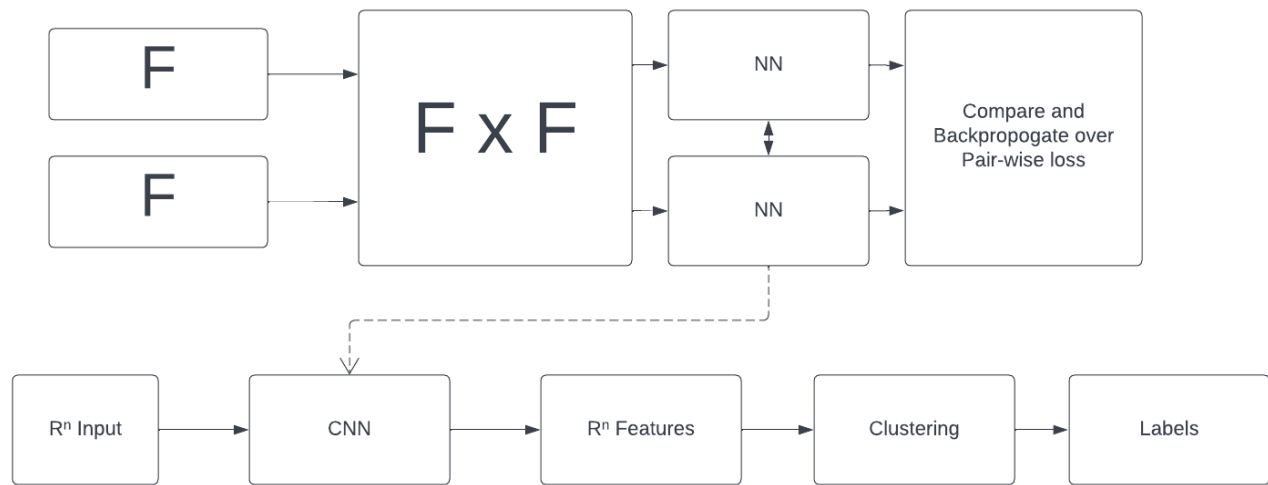


Jason Zhao Question 2 Write up:

The large scale design of this Re-identification algorithm is as follows:



Pairs of detentions are passed into a Siamese Neural Network that consists of two identical NNs that map $R^n \rightarrow R^n$ in parallel connected by a last set of linear layers that compare the difference of the two outputs. The purpose of this is to map the feature space into another feature space that is easier to separate into clusters. The NN is then used to produce an augmented feature space from the detentions that the clustering algorithm can then use to compute clusters. There were many considerations for the clustering algorithm but DBSCAN was chosen due to the how easily it can be modified into a streaming capable variant and the adaptability it offered by not requiring the number of clusters to be known a-priori. However due to the nature of high dimensional data and euclidean distance, K-Means remains a strong alternative given that the number of unique people in the cluster is known beforehand.

The current implementation with DBSCAN requires all detection to be present before clustering is done. However the clustering component of this algorithm is extremely modular and can easily be exchanged for a more streaming friendly clustering algorithm.