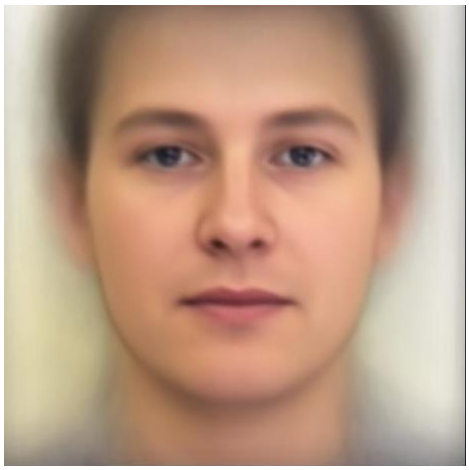
HW4 Report

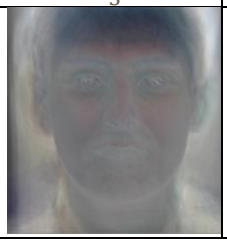
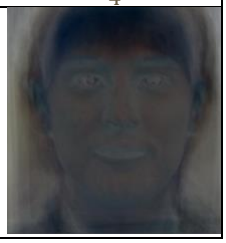
學號：B04502031　系級：電機二　姓名：施力維

A. PCA of colored faces

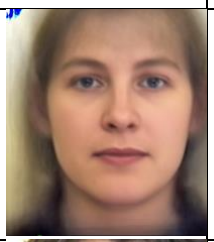
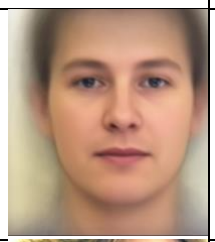
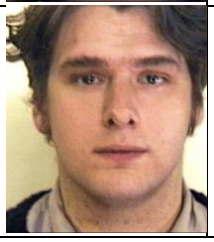
1. (.5%) 請畫出所有臉的平均。



2. (.5%) 請畫出前四個 Eigenfaces，也就是對應到前四大 Eigenvalues 的 Eigenvectors。

λ_1	λ_2	λ_3	λ_4
			

3. (.5%) 請從數據集中挑出任意四個圖片，並用前四大 Eigenfaces 進行 reconstruction，並畫出結果。

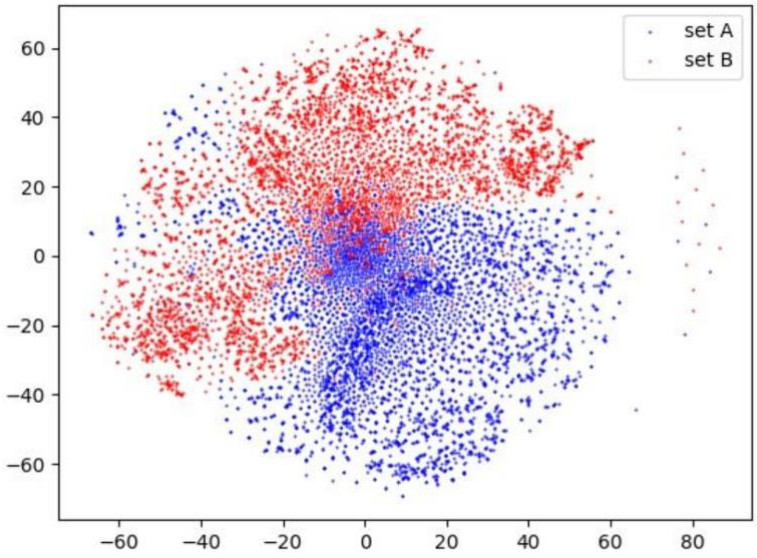
	6	216	196	80
4 個				
原圖				

4. (.5%) 請寫出前四大 Eigenfaces 各自所佔的比重，請用百分比表示並四捨五入到小數點後一位。

λ_1	λ_2	λ_3	λ_4
4.1%	2.9%	2.4%	2.2%

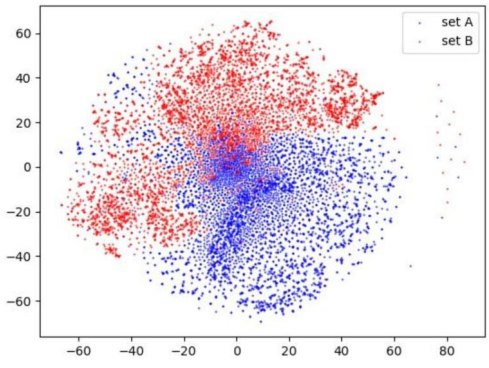
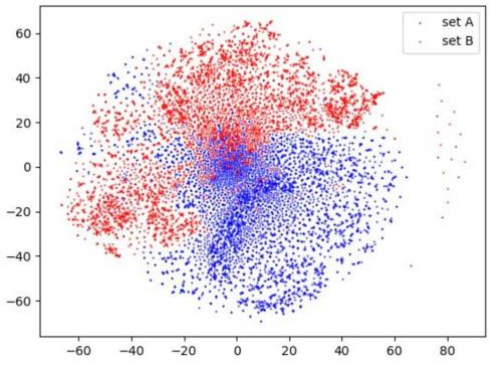
B. Image clustering

1. (.5%) 請比較至少兩種不同的 feature extraction 及其結果。(不同的降維方法或不同的 cluster 方法都可以算是不同的方法)
   兩種方法都使用 PCA 取 400 維來實作，而 cluster 的方法有所不同，其中 PCA 都有開 whitening：
   (a) distance
   做完 PCA 降維之後，根據兩點之間的距離來判斷是否為同一個類別，選取 2000 作為 threshold，做出來的結果為(0.84, 0.83)
   (b) K-means
   K-means 採用的參數為分成 2 類，其他的都是預設參數，最後做出來的結果可以到(1.00, 1.00)的分辨率，不過不同的 random seed 跑出來的結果也會不同，也有可能會分爛。
2. (.5%) 預測 visualization.npy 中的 label，在二維平面上視覺化 label 的分佈。
   採用 100%的 K-means 模型來進行預測，採用 TSNE 降成兩維來做視覺化，畫出來的圖形如下：



可以看出預測出來藍色的部分集中在右下角，而而紅色在左上半部，比較分散一些，中間的部分有些混在一起沒有分開，但整體而言算是有分開來到。

3. (.5%) visualization.npy 中前 5000 個 images 跟後 5000 個 images 來自不同 dataset。請根據這個資訊，在二維平面上視覺化 label 的分佈，接著比較和自己預測的 label 之間有何不同。

Predict	Label
	

從這兩張圖的比對看起來，接近一模一樣，透過 K-means 的分類效果其實很好，只是在 TSNE 上的圖看起來就沒有分到很開，如果要使用 TSNE 來做 clustering，做出來可能就會有一些誤差存在。

C. Ensemble learning

1. (1.5%) 請在 hw1/hw2/hw3 的 task 上擇一實作 ensemble learning，請比較其與未使用 ensemble method 的模型在 public/private score 的表現並詳細說明你實作的方法。（所有跟

ensemble learning 有關的方法都可以，不需要像 hw3 的要求硬塞到同一個 model 中)

本題使用 HW3 來做 Ensemble，取 Public 表現最好的 3 個 model，分別是(public, private) = (0.690, 0.688)、(0.693, 0.675)、(0.700, 0.686)，實作方法是將 3 個 model 的 output 的機率直接平均再挑選機率最大的類別，ensemble 的效果是進步到(0.720, 0.703)。

可以觀察到，由於 model 的選擇是根據 public 來挑選，因此和起來也會是偏向 public 比較高而 private 較低，這也是因為這邊只有挑選 3 個 model，而此種挑選方法可能會造成 bias，有點 overfit public 的 data set。