視覺與語言之關聯於影像理解應用

將圖片中出現的路人甲消除並補足背景畫面已經是十分常見的圖片修飾工具能夠做到的，但是將缺少某些畫面的圖片用其他背景元素擴充整張圖片變成目前人工智慧影像處理的一個重大的挑戰。首先要讓深度學習模型清晰且準確的辨識一張圖片內有哪些元素，並且用一段簡潔有力的句子來描述，能夠準確讀取並辨識才能真正運用在圖片修改或擴充上。對於電腦而已，要學習每個影像需要將元素確實編碼成 0 跟 1，還需要足夠的訓練資料集和標籤輔助才能夠有效率的製作出深度學習模型，我認為這個才是真正困難和繁瑣的過程，因此我認為需要作出一個良好的影像辨識模型，也許需要多個深度學習模型輔助做好預處理。
經過教授帶領研究團隊嘗試各種演算法及資料結構，再搭配 NVIDIA 顯示卡的硬體算力加速，能夠輕易的將一張殘缺的圖片透過深度學習的方式從零到有將圖片重新繪製，我認為將既有的元素和各式各樣考量產生圖片，比重新製作一張圖片，類似 DeepFake 或是虛擬場景更難達成，要如何經由深度學習模型延伸出完美無瑕的圖片，甚至是聲音及圖片描述，不是簡單的輸入輸出資料搭配簡易的隱層層就能夠完成的，希望自己在深度學習的課程中，實作多個競賽題材跟反覆的練習後能夠有些想法跟做法達到這種超常的非監督式學習。