國立清華大學 電機工程學系

實作專題研究成果摘要

Match All Gradient Pixels

梯度像素全匹配

專題領域：資工領域

組　　別：B341

指導教授：孫民教授

組員姓名：謝宗翰

研究期間：　112年2月8日 至 112年11月29日止，共 9 個月

1. 報告摘要

相較於傳統之數學幾何，使用監督式學習之深度學習模型進行點匹配研究可以達到更好得即時效果。然而，若採用現實場景影像進行訓練，會需要消耗大量的人力資源與時間進行所有點匹配的標點作業，因此目前大多數模型採用人工合成影像，或只採取少數幾個影像特徵點進行匹配，缺少許多數據量。因此，本實作專題針對現有的可能實驗點匹配的各種方式，包含光流預測(Optical flow)、相機姿態預測(Camera pose)、特徵點與特徵描述子(Key point and Descriptor)和仿射變換(Affine transformation)等等，進行實作與分析。我們目標是設計出一個監督式深度學習模型，匹配現實物體邊線上的所有像素點，因為邊緣通常擁有較多特徵訊息，且大量的匹配點資訊理論能訓練出更好的模型。為了達成此目的，我們選取了清大校園的真實照片進行仿射變換再做梯度轉換(Gradient)得到梯度影像和匹配點對，以此資料集去訓練深度學習模型。本實作專題針對現有的深度學習模型 Capsnet 進行微調，將 Loss function 從 Camera pose 與預測點的距離，改為真實點與預測點的 L2歐式距離（Mean Square Error），並將資料集改為校園仿射變換影像對進行訓練，在 Mean Matching Accuracy(MMA)@10 pixels 達到良好的效果。仿射變換讓我們能使用真實影像搭配更大量的點進行匹配訓練，但仿射變換後的影像仍為非現實僅是原始圖片得扭曲與平移。因此，未來我們希望設計光流資料集進行跨幀的光流疊加，達到所有點自動匹配且不受 baseline 影響的全新資料集，並使用這大量的資料集去訓練模型達到更好的現實場景匹配效果，減少只有少量特徵點匹配資料量不足與訓練資料集非真實世界影像之缺陷。

2. 報告內容

2.1. Motivation

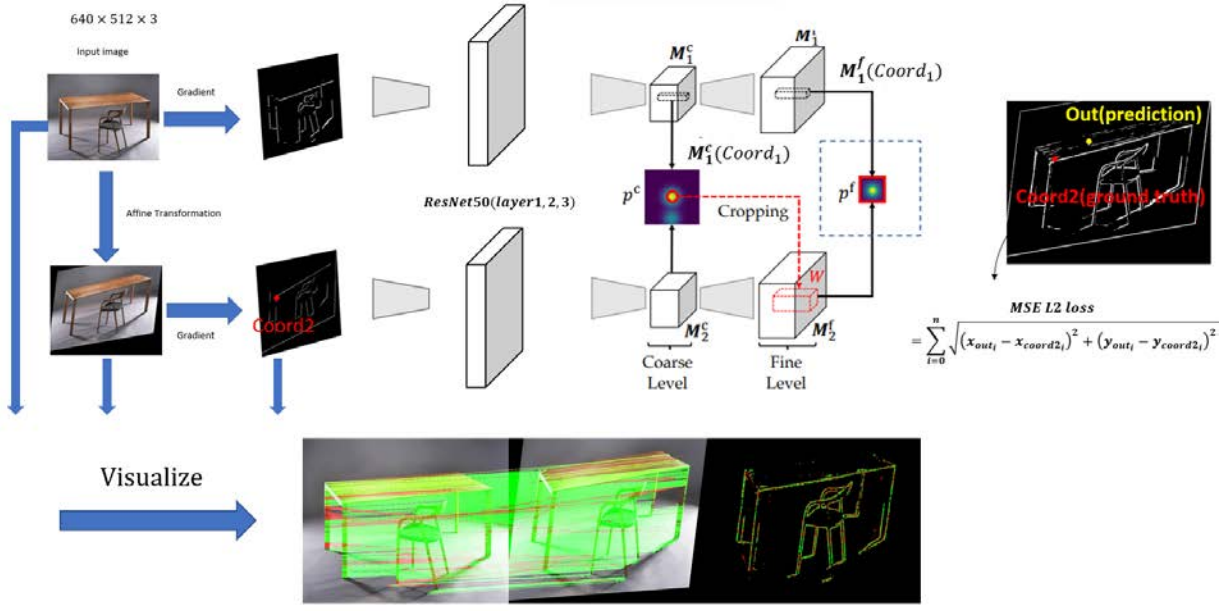
在電腦視覺領域中，點匹配是個經典且熱門的題目，因為在多個領域上有廣泛的應用。例如透過不同場景的特徵匹配，可以辨識物體在不同場景的位置、姿態和運動，在機器人或是自動駕駛，物體辨識和追蹤是個非常重要的功能。此外，若能知道不同角度影像的像素點配對關係，系統可以將影像無縫地拼接在一起得到更完整的圖像，甚至可以估算物體與場景的三維結構（Structure from Motion）對於現在熱門的元宇宙或是空間建築設計有非常重要的影響。現實生活種有很多不同角度不同場景的照片，但是為了訓練點匹配之深度學習模型之前，所需要使用的資料集將消耗大量的人力與時間去進行像素點的配對。因此，目前大多數模型採用合成的簡單幾何影像資料集，或是只採取少數幾個影像特徵點進行匹配，缺少許多數據量，但我們預估若能增加現實生活影像之匹配點數量應會得到更好的效果。

2.2. Purpose

本專題之實作目的在於避免消耗大量人力與時間標記，而擁有大量現實影像的點匹配資料集，以進行點匹配之深度學習模型訓練。先將影像經過梯度轉換再做仿射變換再進微調後的深度學習模型 Capsnet，將 Loss function 改為真實點與預測點的 L2歐式距離（Mean Square Error），並將資料集改為校園仿射變換影像對進行訓練，最後使用 Mean Matching Accuracy(MMA)在10, 9, 8, 7, 6, 5, 4, 3和2pixels 的誤差範圍內做估算。此外，也嘗試使用 Pyrender 實作相機姿態的資料集、使用 OpenCV 的 Lucas Kanade Optical Flow Algorithm 計算光流並進行光流疊加的實驗與概念設計。

2.3. Method

2.3.1. System Design



（圖一）整體研究架構

先將圖片統一大小為寬度(Width) = 640, 高度(Heigh) = 512 和 RGB 三通道(Channel) = 3。其次，將 Image 做 Affine ransform 並記錄所有匹配點座標，再使用 Sobel model 將此 Image pair 做梯度轉換(Filter kernel = 3, Threshold = 0.2)。再來通過 CapsNet 的架構，前面三層 ResNet50做資料與特徵提取，再分別經過 Coarse Level 和 Fine Level Feature Map 預測特徵點的分布期望值得到預測點，使用我們預先準備好的 Ground truth 匹配點(Coord1, Coord2)，透過 Mean Square Error (MSE) L2 loss function 計算預測點預測值與預測點實際值之間的均方誤差，並使用 ADAM 的優化算法更新模型權重參數。最後，設計連線與點雲顏色與分布的方式，使匹配成果視覺化。

2.4. Dataset Preparation

2.4.1. Gradient Pixels

我們使用 Sobel model 以 Filter kernel = 3進行卷積(Convolution)計算，得到水平邊緣與垂直邊緣後，再進行疊加得到完整的邊緣灰度影像(Threshold = 0.2)。使用梯度處理後的影像保留所有邊緣特徵，我們已知重要的特徵點大多出現在邊緣，而直接涵蓋所有特徵點不僅能保留目標物的輪廓、增加匹配點的資料，同時也能減少計算特徵描述子(Feature descriptors)的時間與資源。

$$G_x = I * \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, G_y = I * \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}, G = \sqrt{G_x^2 + G_y^2}$$

梯度轉換將從原先446464個像素點縮減至50122個像素點，縮減了88.8%。此外，我們將剩下的50122個像素點由上到下按照順序儲存，就能知道點與點之間的對應關係。透過梯度轉換與排列紀錄點匹配資訊，以利後續之點匹配模型有大量的 Ground truth 特徵匹配點資料。

2.4.2. Pyrender

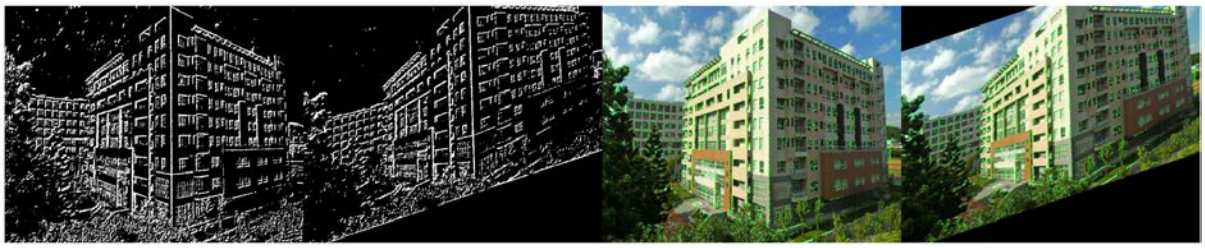
參考論文 Learning Feature Descriptors using Camera Pose，可以直接從相對相機參數來訓練點匹配模型。因此，我們使用 Pyrender 能隨機產生大量3D 場景圖，並記錄所有物體的相機內外參數(Intrinsic and Extrinsic Camera Pose Parameters) 計算得到 Image Pair 的相對相機參數能快速且大量產出不同角度的真實3D 物體影像。然而，效果並不如預期，我們希望能直接使用點對點的 Ground Truth 資料集，並沿用論文的 Network 進行訓練，因此準備了 Affine transformation 的資料集。

2.4.3. Affine Transformation

使用仿射變換(Affine transformation)對影像進行旋轉伸縮和平移，並紀錄變換之矩陣參數 M，得到訓練資料集(Image pair 和一組點對點的 Ground truth)，使用此資料集去訓練進行點匹配的深度學習模型。雖然仍是經由旋轉伸縮平移的資料集，但相對合成的簡單幾何物件，真實圖片與材質更貼近現實場景，又能同時有大量的匹配點 Ground truth 資料以訓練模型。

我們已知在二維圖片上進行線性變換為二維的變換矩陣，然而仿射變換則多增加了平移的變量，因此為了使用仿射變換矩陣，我們需要使用齊次座標，將二維像素點(x, y)轉為三維(x, y, 1)，則仿射變換矩陣應為$M = \begin{bmatrix} R_{11} & R_{12} & t_x \\ R_{21} & R_{22} & t_y \\ 0 & 0 & 1 \end{bmatrix}$，$\begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}$為常見之旋轉、伸縮、平移、推移矩陣，而$t_x$, t_y則為二維平移量。

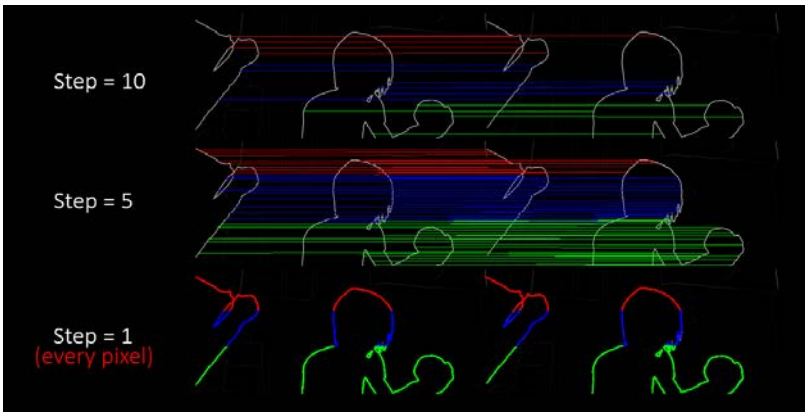
由於我們已經知道仿射變換矩陣 M，我們可以將原始圖片上所有的點投射到變換後的圖片，但這並非我們想要的，我們希望可以匹配所有輪廓邊上的點即可，因此我們將原始圖片先做梯度轉換，濾掉大部分非特徵點，留著這些梯度像素點當作輸入欲預測點(Coord1)。如圖二左邊兩張圖片分別對原始圖片以及變換後圖片進行梯度轉換，並紀錄梯度像素匹配點配對資訊以及兩張影像做為資料集，如圖二右邊兩張完整影像以及上面的綠色對應點。



（圖二）將影像做梯度轉換後濾掉非特徵點。

2.4.4. Optical Flow Accumulation

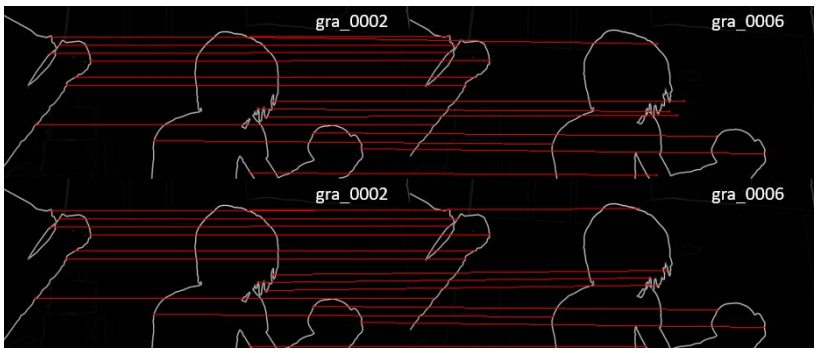
我們使用 OpenCV 內建的 Lucas-Kanade 光流演算法，計算相鄰幀與幀之間的所有位置資訊，嘗試將所有點進行連線並視覺化如下圖三所示，由於需匹配所有像素點，因此選擇先以塗色替代線的方式視覺化，以利視覺化結果是否正確匹配。



（圖三）在 Baseline 很小的情況下完全匹配所有梯度像素。

從圖可見只要幀與幀相鄰不大也就是相機位置 Baseline 不大幾乎可以做到完美的完全匹配，但是如果將 Baseline 拉大會造成像素點的偏移，無法正確匹配兩張圖片上的

所有梯度像素點。不過我們發現，透過光流的疊加可以得到修正，如下圖四所示，從第二幀到第六幀原先會造成一些像素點無法匹配成功，但先透過第二幀到第五幀的光流再加上從第五幀到第六幀的光流，即可修正第二幀到第六幀原先無法匹配的情況，如下圖四所示。



（圖四）光流疊加修正梯度像素點的位置，以得到正確的匹配。

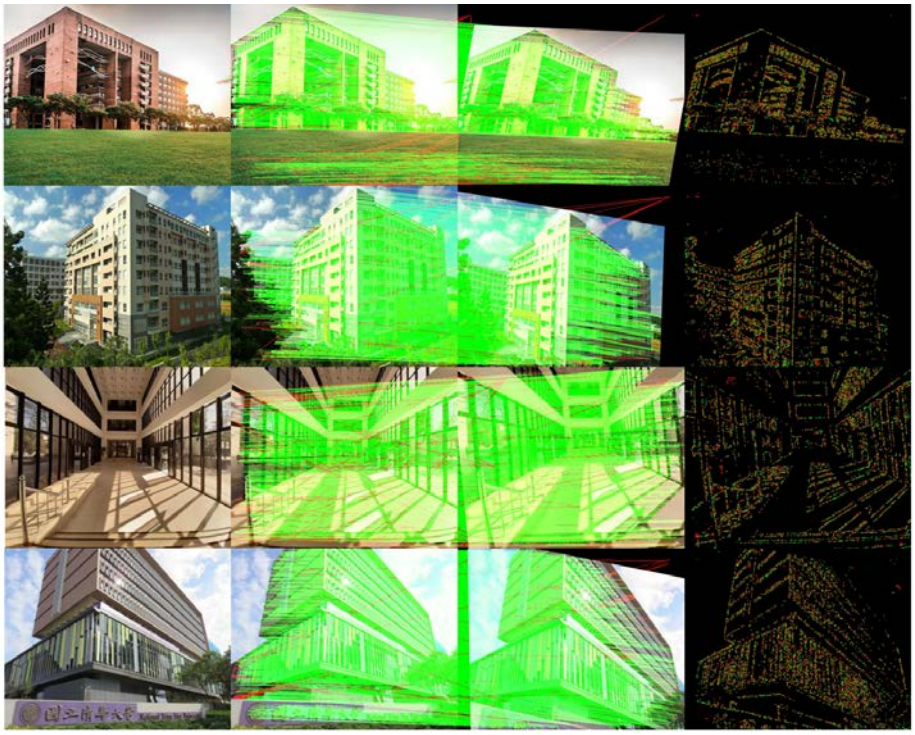
我們從 SuperPoint [1]論文中發現其多為使用簡單幾何圖形得到大量的匹配點後，再使用深度學習模型進行匹配的學習。此研究希望能透過大量的匹配點資訊，得到好的訓練結果，而選用真實圖片進行 Affine transformation，但是變換後終究非真實世界的圖片。因此，希望透過光流疊加的方式，產生大 Baselines 也就是幀與幀跨度很大的影像上所有梯度像素匹配點。以下為理想中希望產出新的資料集的架構流程：

然而，使用光流預測來計算兩張影像之間的點對關係，會涉及到物體遮蔽以及物體遮蔽後又出現的 Lose-tracking 狀況，導致在第一張影像的點無法找到適合的匹配點。由於涉及過多因素與困難，後來僅改用 Affine transformation 生成資料集，並以此資料集去訓練匹配模型達到良好效果。而針對 Lose-tracking 的狀況，我在2023ICCV 最佳學生論文獎 Tracking Everything Everywhere All at Once [2]看到用以預測全局一致與影片運動方向的光流優化方法，也許未來可以結合產生資料集。

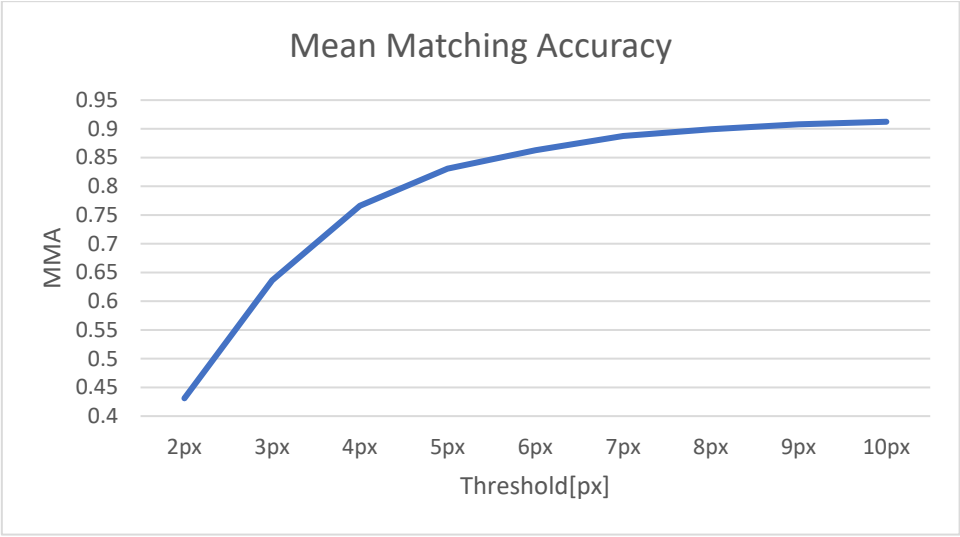
2.5. Model design

採用論文 Capsnet [3]的架構：前五層使用 ResNet-50 [4]搭配 ImageNet-pretrained weights 初始化參數讓收斂加速，之後加上額外的卷積層 Coarse-level feature map 和 Fine-level feature map。不同的是，我們平均每張圖片輸入6000個 points 做點匹配訓練，使用 Adam 優化算法在 Learning rate 為0.0001去訓練模型，並直接把 Loss function 原先為預測點與極線的距離L_{ep}，改為預測點與 Ground truth point 的歐式距離(L2 distance)。此外，論文 [3]中使用 ResNet50 搭配 ImageNet-pretrained weight 去訓練模型，ResNet (Residual Network) 最經典的優勢是深度非常深，且又不會因為模型網路太深產生 Degradation problem，因為有一個類似短路的路徑 Shortcut connection，可以使殘差值為零沒學到東西輸入等於輸出時，僅透過複製上層的特性做 identity mapping。ResNet50網路較深的特性也讓影像辨識上有很好的效果，因此選擇採用此網路作為架構的一部份，並在最後針對特定匹配問題加入額外的卷積層 Coarse-level feature map 和 Fine-level feature map。

2.6. Results



（圖五）最左邊為原始圖片，中間為原始圖片與變換後圖片的梯度像素匹配預測結果，綠色線為正確匹配，紅色線為匹配錯誤。最右邊為點的散佈圖，綠色點為實際值，紅色點為預測值。



（表一）不同門檻下匹配的準確度趨勢圖。

	Threshold								
Pixels(px)	**@10px**	**@9px**	**@8px**	**@7px**	**@6px**	**@5px**	**@4px**	**@3px**	**@2px**
Accuracy	**0.91233**	**0.90783**	**0.89917**	**0.8875**	**0.86283**	**0.83083**	**0.766**	**0.63617**	**0.4312**

（表二）不同門檻下匹配的準確度數據表

此實驗結果已達成專題之實作目的，透過更改模型Dataset以及Loss function之設計，在平均10 pixels以內的誤差準確度MMA@10px達 0.91，在 3 pixels內也有超過一半以上的匹配正確率。也就是說，我們欲觀測的6000個梯度像素點中，有54600個點投射在正確匹配點附近不超過10個像素距離，已足夠定位目標以及目標輪廓。然而，在光流疊加的資料集因遮蔽與Lost-tracking的問題未能解決，因此僅完成Affine transformation的設計。

2.7. Conclusion

我們最終使用仿射變換的校園場景資料集2000張在現有的深度學習模型 Capsnet 上進行訓練，將 Loss function 改為真實點與預測點的 L2歐式距離，在 Mean Matching Accuracy(MMA)@10 pixels 得到良好效果。仿射變換讓我們能使用真實影像搭配更大量的點進行匹配訓練，但仿射變換後的影像仍為非現實僅是原始圖片得扭曲與平移。因此，未來我們希望設計光流資料集進行跨幀的光流疊加，達到所有點自動匹配且不受 baseline 影響的全新資料集，並使用這大量的資料集去訓練模型達到更好的現實場景匹配效果，減少只有少量特徵點匹配資料量不足與訓練資料集非真實世界影像之缺陷。

3. Reference

[1] D. DeTone, T. Malisiewicz and A. Rabinovich, "SuperPoint: Self-Supervised Interest Point Detection and Description," *Computer Vision and Pattern Recognition*, pp. 337-33712, 2018.

[2] Q. Wang, Y.-Y. Chang, R. Cai, Z. Li, B. Hariharan, A. Holynski and N. Snavely, "Tracking Everything Everywhere All at Once," *ICCV* , 2023.

[3] Q. Wang, X. Zhou, B. Hariharan and N. Snavely, "Learning Feature Descriptors using Camera Pose Supervision," *ECCV*, 2020.

[4] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *Computer Vision and Pattern Recognition*, 2016.

[5] 清大校友會, "清大校友中心校友總會," [Online]. Available: https://alumni.site.nthu.edu.tw/p/404-1346-115103.php?Lang=zh-tw.

[6] D. J. Butler, J. Wulff, G. B. Stanley and M. J. Black, "A naturalistic open source movie for optical flow evaluation," *European Conf. on Computer Vision (ECCV)*, pp. 611--625, oct 2012.

4. 心得感想

在這次的專題實作中，其實我學到最多且感受最大的並不是電腦視覺相關的深入知識，而是如何在有限的研究時間內，判斷什麼知識是你需要的？什麼是你不需要的？並把時間投入在需要的知識上深入研究與學習，而避開不需要的知識。我曾經花了大把時間在研究 SLAM 的數學計算、Key point 的描述子、OpenCV 內建的 LK 光流計算、Floor Plan estimation、Layout estimation，甚至是學習怎麼去訓練一個最簡單的深度學習模型，因為我一直覺得我需要把軟體背景知識補齊，且關於深度學習電腦視覺領域的論文非常多且更新率快，除了熟悉背景知識以外，也時常需要關注最新的論文並同時參與實作，過去常常在不知不覺中偏離了主題而不自知。好學是一件好事，但不是每件事都需要知道才能開始做研究和實驗，學你所需，需要的時候再學，我覺得也是在這事專題中又或者是未來的研究中應該邁向的學習模式，學無止盡我們不可能把所有知識都學完，就算有那一天大概也沒有時間做研究。Learning by doing，從憑空生出自己大量的資料集，到自己設計模型的輸入、輸出與網路架構，一整套流程雖然還沒有完成，但我學到很多也學以致用。在實習過程中因為學了 OpenCV 且有 python 自動化與檔案管理的概念，讓我在實習過程中發明了一種熵的視覺化方法並先有了論文，算是額外岔出的一個小確幸，期待後續在視覺的研究不管是在軟體演算法的開發上或是硬體資源配置的加速上，都能有更多的突破。