應用人工智慧分辨桌球運動員之熟練度

專題生：呂浩凱、孫唯真、郭育誠

指導教授：林政宏

摘要—隨著科技的進步，人工智慧的應用也越來越廣泛和複雜，深度學習技術在近幾年的快速發展，使其成為解決複雜問題的重要工具，從最初圖像識別到近期非常熱門的動作辨識，深度學習功不可沒。我們的研究目標是建立一套能辨識桌球運動員揮拍動作的系統，以此評估運動員的技術水平，並將其分為初學和精熟兩種熟練度等級，期望最終的專題成果，能對桌球的訓練及評估帶來客觀且實質的貢獻。本專題主要引用 Temporal Score Network，結合卷積神經網絡(CNN)和殘差神經網路(Resnet)的架構，以我們自行開發的桌球動作資料集(Table Tennis Pose Dataset, TT Pose Dataset)進行訓練與測試。隨後，我們借鑒了Grad-CAM技術，將其應用於原始架構，並生成模型分類準則的熱區圖，以提供深度學習黑盒子的可視化解釋。此外，我們開發了一個簡單方便的使用者介面(UI)，讓用戶能夠自行上傳桌球揮拍動作影片，影片通過神經網路生成Grad-CAM結果和其分類基準的具體部位資訊後，再由UI同步播放原輸入影片和Grad-CAM影片，從而提供使用者更直觀快速的分析結果。

***關鍵字*—深度學習(Deep learning)、影像辨識(Image recognition)、卷積神經網絡(Convolution Neural Network, CNN)、殘差神經網路(Residual Network, ResNet)、梯度加權類別活化映射(Gradient-weighted Class Activation Mapping, Grad-CAM)、使用者界面(User Interface, UI)、桌球動作資料集(Table Tennis Pose Dataset, TT Pose Dataset)。**

I. 導論

近年來，深度學習領域的相關技術以驚人的速度蓬勃發展，使其成為了眾多研究者解決複雜問題的理想選擇，其應用在多個科研領域如計算機視覺、自然語言處理和語音識別等，已經有重大貢獻且取得了顯著成果。深度學習的優勢在於能夠從大量數據中，模擬人體神經網路的複雜結構，自動提取特徵並進行高效的學習。

有賴於卷積神經網路(CNN)的發展，動作辨識在近幾年也有許多突破性的研究成果。然而，在運動項目的應用上，由於其包含複雜的身體運動和協調動作，目前的研究發展主要依舊集中在辨識粗粒度的動作上，傳統的分析方法往往難以捕捉到運動員的微妙動作和技術細節。但在某些情境下，分辨動作的細微差異才是主要訴求，例如不同的桌球運動員在動作上也許只有一點差異，卻會造成擊球和學習成效的巨大差異。有鑑於桌球是一項在少數球類運動中較易於穩定追蹤運動員動作的運動項目，且我們的研究團隊中有一位專題生對桌球運動懷有極大熱忱，我們期望能藉此專題實作的機會，建立一個能夠讓電腦自動識別桌球運動員揮拍動作的學習系統，並進一步對動作熟練度進行評估和分類。

本專題旨在運用現有的深度學習技術，並結合我們自行開發的桌球動作資料集(TTPose Dataset)，建立一套桌球運動員的揮拍動作辨識系統。透過此系統，我們能夠快速且客觀地評估桌球運動員的技術水平，將其分類為初學和精熟兩種熟練度等級，並在動作表現上提供有價值的參考結果。研究過程主要採用Temporal Score Network[1][2]，以CNN和Resnet的架構作為基礎[3]，建立深度學習模型，再套用我們自行開發的桌球動作資料集。此資料集包含桌球初學者與精熟者的揮拍動作短影片各600部，並平均分配在訓練集與測試集當中。實作過程會先以900部的訓練影片進行模型的訓練與優化，再以300部的測試影片測試模型的穩定性與可用性。實驗結果顯示，我們在測試集的分類準確率最高可達92.377%，證明此模型有一定能力可以成功分辨桌球初學者和精熟者。

在確定模型有足夠能力對陌生影片進行辨識和判斷後，為了能在桌球的訓練、技術提升和競技水平評估等領域，對運動員、教練或裁判提供客觀參考和實質貢獻，我們參考了Grad-CAM的理論基礎及應用[4][5]，並將其套用在原基礎架構上，藉由觀察在運動員揮拍動作上的熱區圖，對模型分類的標準提供一個可視化的解釋。此外，我們開發了一個高度使用者友好且易於操作的使用者介面(User Interface, UI)，讓用戶能夠自行上傳欲進行辨識之桌球揮

拍動作影片，並快速獲得該運動員的精熟度分類結果，以及其桌球關鍵動作的具體部位資訊。此界面的開發能大幅提升用戶體驗，並提供更直觀快速的分析結果。

本專題的主要貢獻包括：

I. 改良優化Temporal Score Network的模型架構，並將之套用在桌球運動的動作辨識和精熟度分類上。分類準確率最高可達92.377%。

II. 自行建構一個全新的桌球動作資料集(Table Tennis Pose Dataset, TT Pose Dataset)，包含900部訓練與300部測試、共1200部桌球揮拍短影片，透過此資料集進行模型的訓練與測試。

III. 開發使用者介面(UI)，提供用戶更簡單方便的使用感受和直觀快速的分析結果。

此篇論文的結構如下：第一段會介紹深度學習的背景和發展情況，並概述我們的研究目標和方法。第二段回顧了早期的研究工作，並簡單介紹相關的技術內容。第三段會詳細介紹我們的整體的研究方法及架構，並說明其背後的技術細節。第四段會報告我們在數據集上的實驗結果和性能分析。最後，第五段總結了我們的研究成果，並提出未來的研究方向。

II. 文獻探討

A. *Residual Network (ResNet)*

殘差神經網路(ResNet)是一種使用殘差塊(Residual block)的卷積神經網路，能夠進一步加深模型的深度，並提升其泛化能力以及圖像識別的精度。其最主要的特點是引入了殘差學習(Residual Learning)的概念，即每個殘差塊都學習對輸入特徵進行修改，從而使輸出更接近目標。殘差神經網路能夠使模型有效地學習較小的殘差，進而獲得更好的泛化能力。

從技術層面來說，每個殘差塊可以直接跳過一個或多個層，解決梯度消失和梯度爆炸等問題。此外，ResNet使用了批量標準化(Batch Normalization)技術來加速模型訓練，並通過權重衰減(Weight Decay)和Dropout，控制模型的複雜度和防止過度擬合(overfitting)[3]。

B. *ViTPose*

ViTPose是一種基於Transformer的人體姿態估計方法。此方法引用了一種全新的神經網絡架構——Vision Transformer (ViT)，和傳統CNN不同的是，ViT能以較簡單直觀的架構和較少的參數，有效的進行特徵擷取並學習。

ViTPose模型使用Transformer中的自注意力機制來捕捉姿勢關鍵點之間的依賴關係，並通過多層感知器來預測每個關鍵點的位置。此外，ViTPose還引入了一種新的位置編碼技術，使得模型能夠有效處理不同大小和比例的輸入圖像，進而提高模型的泛化能力和穩定性[7]。

C. *Action Quality Analysis (AQA)*

由於深度學習技術在動作識別(Recognizing actions)領域已經有很大的進展，越來越多學者將研究方向聚焦在動作評估(Assessing actions)上，動作質量評估在醫療保健、運動和視頻檢索等實際應用中具有許多潛在價值。迴歸模型需要評估動作質量時，需要從影片框架中擷取兩組特徵，分別是低階特徵和高階特徵，進行模型的訓練和評分。低階特徵是一種層次特徵，透過捕捉邊緣和速度的時空濾波器進行擷取；高階特徵則是提供如何改善動作的反饋，相比低階特徵直接捕捉像素的信息，通常難以解釋，高階特徵具有可解釋性。在電腦視覺研究中，儘管動作質量評估具有相當潛力，此技術仍然有很大的發展空間，才能與專家評分的結果相媲美[6]。

D. *TTNet*

TTNet是一個輕量級多任務的神經網路架構，可用於實時處理高分辨率的桌球影片，提供像素級準確度的時間和空間的數據，包括桌球發球、彈跳和觸網等，替自動裁判系統的推理得分提供核心的信息。此研究的關鍵技術在於透過連續的幀，而非單一幀來捕捉桌球運動信息，並排除其他可能被誤判的雜訊。這項實時多任務的深度學習應用，可以在桌球比賽過程蒐集額外的資訊，提供裁判在決策時的參考依據[8]。

E. *Grad-CAM*

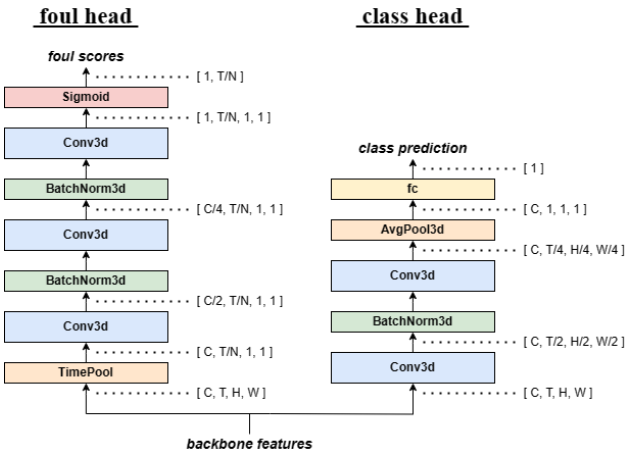
目前深度學習在圖像分類、檢測和語義分割等任務中表現出色，但深度學習模型通常是個黑盒子，難以解釋其預測結果的基礎原因。為了解決這個問題，越來越多的學者開始研究如何解釋深度學習模型，Grad-CAM就是其中一種，它能夠在不需要重新訓練模型的情況下，通過定位模型決策的重要區域，並且基於這些梯度生成視覺化的解釋，使我們能夠更好地理解深度學習模型的行為。

具體來說，Grad-CAM使用了CNN中的最後一層卷積層，或是全局平均池化層(Global Average Pooling, GAP)，將其視為模型的特徵提取器。通過計算這些特徵對輸出類別的貢獻，Grad-CAM可以生成一個與特定類別相關聯的視覺化熱區圖，清晰地顯示出對這個類別預測最有貢獻的圖像區域，從而有效解釋了模型預測的基準。Grad-CAM的好處是，它不需要任何額外的標注信息或模型訓練，並且可以應用於各種深度學習模型，包括卷積神經網絡(CNN)、遞歸神經網絡(RNN)和類神經網絡(GAN)等[4][5]。

III. 原理與架構簡介

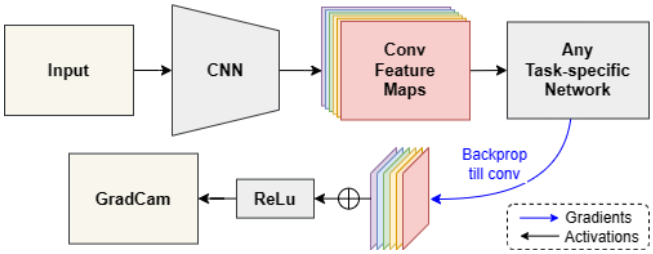
本專題參考論文A Temporal Scores Network for Basketball Foul Classification 所提出的架構 [1][2]，期望透過深度學習辨別桌球初學與精熟者的動作差異。論文中提出的Temporal Scores Network（時空分數網路），是一套細粒度的(Fine-grained)動作識別網路，此網路可以套用於許多現有的神經網路架構，包括 3D-Resnet50、3D-wide-Resnet50、R(2 +1) D-Resnet50 和 I3D-50。藉由取代掉原本CNN的最後一層，可以增進原先模型探測細粒度動作的能力，並加強它們辨識細微動作的準確性。

時空分數網路的架構包含兩個端點(heads)，如圖一所示，分別是左邊的分數端和右邊的類別端。分數端的目標是提取高時間分辨率的特徵，輸入的影片首先會經過第一層時間池(TimePool)，也就是是三維卷積層(3D-Conv)，之後再陸續經過三個三維卷積層後，輸出最終預測的動作評分；而類別端的目標是提取時空整合特徵，影片輸入後會經過兩個三維卷積層、一個平均池化層(average pooling)和一個全連接層(fully connected)，最終得出各類別的機率。



圖一、Temporal Score Network架構示意圖。左邊的分支是分數端、右邊是類別端，維度的格式分別是[時間, 層, 高, 寬]。

為了進一步得知神經網路預測類別的依據，我們將時空分數網路輸出的結果丟進Grad-CAM[4][5]，並查看輸出的熱曲圖進行判斷。Grad-CAM的架構總覽如圖二所示。分數端對輸入圖像進行前向傳播(Forward propagation)後，會得到最後一個卷積層的輸出特徵圖，接著經過反向傳播(Back propagation)，計算目標類別相對於最後一個卷積層的梯度資訊。反向傳播中的梯度資訊會經過全局平均池化(Global Average Pooling, GAP)，將每個特徵圖上所有像素的梯度平均值作為該特徵圖的權重，用於計算特徵圖上每個像素對最終分類結果的貢獻比例。接著，我們將每個特徵圖上的權重與對應的特徵圖相乘，並將它們加權求和，得知網路是根據圖片的哪一區塊進行判斷分類的，最終得到一個熱區圖，顯示不同位置對於最終分類結果的相對重要性，視覺化神經網路的分類結果。



圖二、Grad-CAM架構示意圖。剛開始輸入進一個圖片和一個目標類別，接著通過模型的CNN部分前向傳播圖片，並通過任務導向的網路來計算該類別的原始分數。目標類別的梯度會被設置成1，此梯度隨後被反向傳播到整流卷積特徵圖，最後獲得Grad-CAM藍色熱區圖。

為了讓其他人能更好地運用本專題的研究成果，我們透過PyQt5開發了使用者介面(UI)，希望透過此界面的設計和建構，提供用戶自行上傳影片的功能，並迅速輸出該影片的精熟度分類結果，及其桌球關鍵動作上的熱區影片。我們相信UI的開發能大幅提升用戶的使用感受，並提供更直觀快速的分析結果。在將Grad-CAM導入UI程式碼的過程中，我們有遇到一些難題，包括像是讀取影片時的複雜路徑設定和硬體效能限制等，以及如何將長影片拆分測試後再重組輸出。最終，我們決定將界面設計為自動輸出分類結果，並同步播放測試和Grad-CAM的影片。

IV. 實驗步驟、過程與結果

本專題採用的「桌球動作資料集(TTPose Dataset)」是自行建構的全新資料集，藉由招募桌球志願者，拍攝蒐集79部初學與精熟者的揮拍動作影片，並進行後續的剪輯處理而成。拍攝的桌球動作包括平面膠皮的正手與反手擊球，角度取正面以及左右側各45度，場景是固定不變的，拍攝影片的檔案類型為.MOV。接著，拍攝影片會進行人工剪輯處理，剪輯時長以一次揮拍為單位，目的是為避免模型學習到揮拍以外的因素（如揮擊速度、撿球等），剪輯影片大小為256*256，剪輯影片的檔案類型為.mp4，最終剪了初學和精熟者各600筆，共1200筆影片。在資料集規格方面，影片以3:1的比例分成不重複的訓練用與測試用資料集，分別包含900部及300部，且訓練與測試資料集皆平均分配初學及精熟者各半， label 0代表初學者、1代表精熟者。

輸入的影片在進模型跑之前，會先經過Transform隨機裁切成224*224，加快神經網路的執行速度，並將其標準化(Normalize)到[-1, 1]之間，加速模型的收斂速度，最後經由dataloader，將影片送進時空分數網路模型進行分類預測。

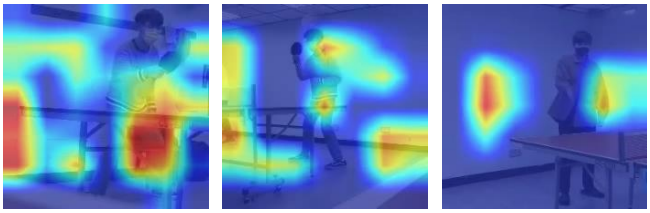
模型訓練後的最佳測試結果如表一所示。由於dataloader中各種參數的數值會很大程度地影響模型訓練的結果，我們在參數設定上琢磨了一陣子，最終選定num_worker為32，batch size為8，epoch為80，可以最佳化模型的訓練效率與結果。

表一 以桌球動作資料集訓練Temporal Score Network後的測試結果。分別是雙端（分數和類別端）和單端（類別端）的最佳測試準確率。

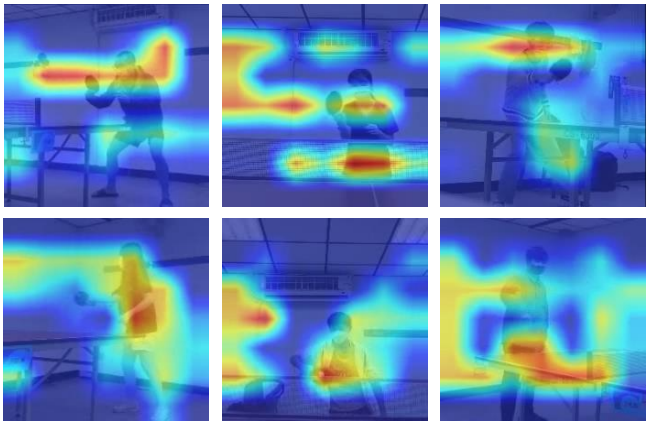
METHOD	TEST ACCURACY
Dual-head network	90.955
Single-head network	92.377

從數值可以看出，模型在大方向上有能力準確分辨初學和精熟者的差別，並且使用單一個類別端比雙端分類模型的測試結果還要準確，單端的最佳測試準確率為92.377，雙端最佳則為90.955。我們分析單端準確率比雙端高的原因，可能是因為雙端的架構較適合應用在長影片的判斷。原先時空分數網路之資料集影片長度約為5~10秒，雙端網路的表現較突出，而我們訓練時之資料集是1~2秒的短影片，所以單端網路訓練測試準確率較高。此外，我們還發現神經網路的分類機率都呈現一面倒的趨勢，顯示模型對於自己的分類結果非常肯定，無論結果是正確或錯誤。我們分析此現象可能是因爲神經網路判斷選手熟練度時，不是依據選手之肢體動作，所以我們決定透過Grad-CAM來進一步了解神經網路辨識之區域。

為了進一步確認模型的判斷依據，我們將每張圖片對應的分類機率丟進Grad-CAM，透過觀察熱區的分佈判斷模型是根據打者的哪個部位進行分類。部分Grad-CAM的輸出結果如圖二、三所示。從圖二可以明顯看出，錯誤分類的Grad-CAM結果確實沒有很理想，熱區分佈的位置很零散也很奇怪。但從正確分類的圖三來看，我們發現精熟者的熱曲主要分佈在手以及身體，這個結果能夠合理解釋模型的判斷依據，反觀初學者的熱區，有些分佈在天花板或桌腳，這讓我們不禁懷疑模型是否能夠精準判斷初學者，而不是透過二元分類的排除法進行分類。

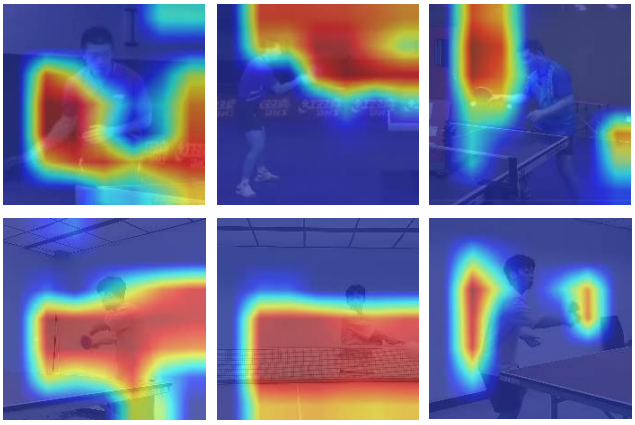


圖二、GradCam視覺化模型分類錯誤的測試集影片的結果。左邊兩張是精熟者，最右邊是初學者。圖中皆為Grad-CAM輸出影片的截圖。



圖三、GradCam視覺化模型分類正確的測試集影片的結果。上方三張是精熟者，下方三張是初學者。圖中皆為Grad-CAM輸出影片的截圖。

為了進一步測試模型的穩定性和準確度，我們從網路上抓取開源的桌球精熟者影片4部，由於開源的初學者影片較少，我們又自行額外拍攝桌球初學者影片6部。接著，將全新的桌球揮拍長影片送進模型測試，並且同樣輸出熱區圖觀察測試結果。神經網路輸出的結果顯示，開源影片的最高測試準確率為84.9%，比原先的測試集稍低一點，但仍具有相當參考性，而分類機率同樣呈現一面倒的趨勢。Grad-CAM的輸出結果如圖四所示。我們推測或許因為自行額外拍攝的影片畫質較好，且與原先資料集相似，初學者的熱區圖結果比開源精熟者來的好。

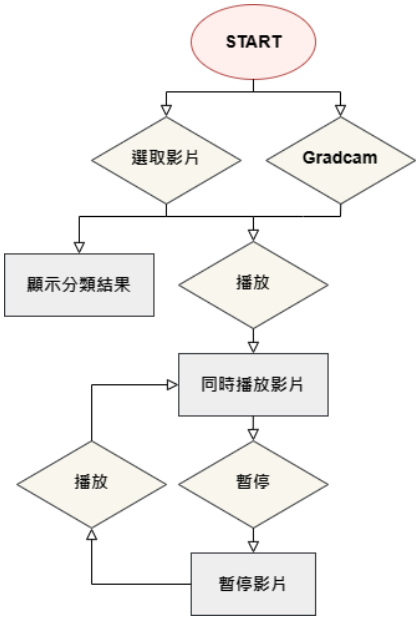


圖四、Grad-CAM視覺化分類正確的開源影片的測試結果。上方三張是分類正確的精熟者，下方三張是分類正確的初學者。圖中皆為Grad-CAM輸出影片的截圖。

由於此專題的最終目標是將成果落地實行，協助桌球運動員、教練或裁判在訓練及參賽過程中，對於桌球的訓練、技術提升和競技水平評估等有更科學的指標，我們決定開發一個使用者界面(UI)，讓用戶——無論是運動員、教練或裁判——都能輕鬆快速地得知該桌球揮拍動作的精熟度分類結果，以及其關鍵動作的具體部位資訊。其界面設計和操作流程圖可以參考圖五、圖六。

圖五、使用者界面的外觀及功能設計。

使用者進入界面後，首先會看到四個按鈕，分別是選取影片、Grad-CAM、播放和暫停。用戶可以點按選取影片和Grad-CAM按鈕，上傳欲進行辨識之桌球揮拍動作影片，接著點選播放按鈕，界面上便會開始同步播放原影片（左側）和相應的動作熱區影片（右側），並輸出該運動員被分類為精熟者亦或是初學者，用戶也可以選按暫停和播放的按鈕，同時控制兩部影片的顯示狀態。我們相信，此使用者界面的開發能大幅提升用戶體驗，提供更直觀快速的分析結果與簡單方便的操作流程。



圖六、使用者界面(UI)的操作流程圖。界面上包含四個按鈕，分別是選取影片、Grad-CAM、播放和暫停。

V. 結論與心得

在本文中，我們提出和優化了一種能辨識桌球揮拍動作，並分類其精熟度的深度學習架構，透過自行蒐集建構的桌球動作資料集(TTPose Dataset)進行模型的訓練和測試。研究結果表明，我們的模型在桌球運動員的精熟度分類，具有良好的成效及參考價值。再者，我們開發了一個使用者界面，讓用戶能以更簡單直觀的方式自行操作界面，快速獲得該輸入影片資訊。未來希望可以拍攝更多角度的資料集，使神經網路有能力分辨多方位的動作，也期待能再精進模型篩選判斷關鍵動作的能力，並增進UI的功能性與完整性。

很開心能藉由此專題實作的機會，對深度學習領域有更多的了解及應用。從一開始的論文學習到後續的實作過程，我們了解到深度學習的博大精深，及其廣泛且強大的應用。期望本專題的研究成果，能對桌球的訓練及評估提供實質客觀的參考。

誌謝

感謝指導教授和實驗室學長們的指引和教導，讓我們在過程中有明確的研究目標和努力方向，也給我們很多收集整理資料集以及撰寫篇論文的建議與協助。

參考資料

[1] P. Y. Chou, C. H. Lin, W. C. Kao, Y. F. Lee, and C. C. Hsu, "A Temporal Scores Network for Basketball Foul Classification," *in 2022 IEEE 12th International Conference on Consumer Electronics (ICCE-Berlin)*, 2022.

[2] C. H. Lin, M. Y. Tsai, and P. Y. Chou, "A lightweight fine-grained action recognition network for basketball foul detection," *in 2021 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, 2021.

[3] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition" *in Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[4] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," *in Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[5] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *in Proceedings of the IEEE international conference on computer vision*, 2017.

[6] H. Pirsiavash, C. Vondrick and A. Torralba, "Assessing the Quality of Actions," *in Computer Vision – ECCV 2014*, 2014

[7] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "Vitpose: Simple vision transformer baselines for human pose estimation," *in arXiv preprint arXiv:2204.12484*, 2022.

[8] R. Voeikov, N. Falaleev, R. Baikulov, "TTNet: Real-time temporal and spatial video analysis of table tennis," *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

指導教授（親簽）：