2025 圖形識別期末報告

題目：透過視覺語言模型與傳統物件追蹤演算法追蹤物件

繳交日期: 2025 年 6 月 5 日中午 12 時以前。

系所:電控工程研究所 **級別:** 碩一

姓名: 朱彥勳 **學號:**

一、 主題描述：

在電腦視覺領域，物件追蹤是一項被持續研究的主題，傳統的物件追蹤系統通常需要透過手動選擇或事先標記目標物，這種方法在使用上並不直覺，也限制了即時性與便利性。而隨著目前視覺語言模型(Vision-Language Model, VLM)的發展，現在視覺語言模型能夠有效的辨識出目前畫面中的物體並且找出其位置，但目前所需的計算量依舊龐大，需要花費大量運算資源，無法達到即時的預測。因此，本次專題的主要目的是透過結合視覺語言模型與傳統快速物件追蹤演算法，實現一種直覺且高效的多物件追蹤系統。透過自然語言輸入來定位目標，使用者無須事先框選，系統即能自動定位並追蹤指定的物件，極大地提升了使用的便利性與普及性。

二、 方法描述：

本專題的所有程式都在 Google Colab 平台進行開發，使用 T4-GPU 進行運算。程式碼連結。專案 Github。系統架構首先利用了目前開源且強大的視覺語言模型 Qwen-2.5-VL 進行物件識別與初始定位。使用者只需輸入如「追蹤 AirPods」、「追蹤耳機」的自然語言提示，模型便能在影片的第一影格中精確地產生該物品的邊界框座標。接下來，系統將這個初始邊界框座標傳遞給 OpenCV 所提供的傳統物件追蹤器，我選用 CSRT 演算法，進行後續即時追蹤。

以下針對技術細節進行介紹:

1. Qwen2.5 VL:

Qwen2.5 VL(千問 2.5)是由是阿里雲 Qwen 團隊在 2025Q1 發布的最新視覺-語言大型模型（LVLM），主打原生高解析度處理、長影片理解與精確空間-時間定位的能力，同時保留 Qwen 2.5 模型在程式推理與多語言對話上的水準。

在空間維度中，Qwen2.5-VL 能夠動態的將不同尺寸的圖像轉換為不同長度的 Token，還直接使用圖像的實際尺寸來表示檢測框和點座標，而不進行傳統的座標歸一化，詳細技術細節可參考這篇網站。

目前在主流的視覺模型中 Qwen2.5-VL 在多個 Benchmark 都勝過其他閉源的模型，這也是我選擇他的原因。

由於計算資源有限，我選擇 **Qwen 2.5-VL-3B-Instruct** 作為我的推論模型。

名稱	參數量	FP16 VRAM（推論）	特色
Qwen 2.5-VL-3B-Instruct	≈ 3.1 B	~15 GB	邊緣部署、LoRA 微調首選。
Qwen 2.5-VL-7B-Instruct	≈ 7.0 B	~23 GB	性能／成本平衡，社群教學常用。
Qwen 2.5-VL-72B-Instruct	≈ 72 B	8×80 GB A100	逼近 GPT-4o/Claude 3.5 在多模態基準的旗艦版。

表(一) 模型尺寸與運行需求

2. 物件識別演算法:

在 OpenCV 中常見的物件演算法主要有 KCF、MOSSE 與 CSRT 等，我選擇使用效果較好的 CSRT 驗算法，追蹤器最主要適用於判斷前後 Frame 的物件特徵，並持續判斷物件位置。

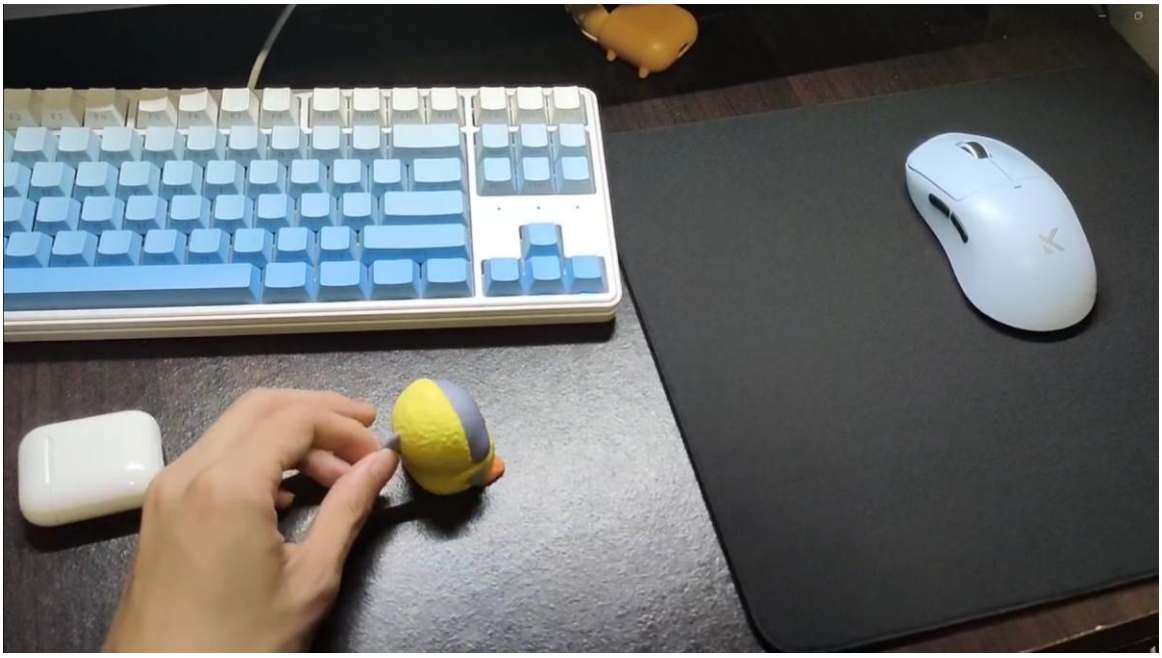
CSRT 介紹:

CSRT（Channel and Spatial Reliability Tracker）是 CSR-DCF [“Discriminative Correlation Filter with Channel and Spatial Reliability”] 的 OpenCV 改寫版本，原始論文在 CVPR 2017 發表，屬於 DCF（Discriminative Correlation Filter）系列 的短期單目標追蹤器。CSRT 透過「通道可靠度」與「空間可靠度遮罩」兩個機制，解決傳統 DCF 對邊界循環假設過度敏感、容易漂移到背景，以及某些通道（特徵）表現差卻仍被平均採用的問題。具體介紹可以參考此篇網站。

三、 結論與自評：

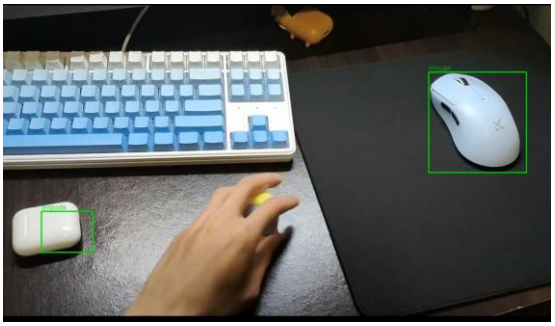
結果展示:

為了展現 VLM 的辨識物體靈活性，我拍攝了一段影片(可參考 Github 連結)，最主要包含了四樣物體:鍵盤、玩偶、滑鼠、耳機(Airpods)

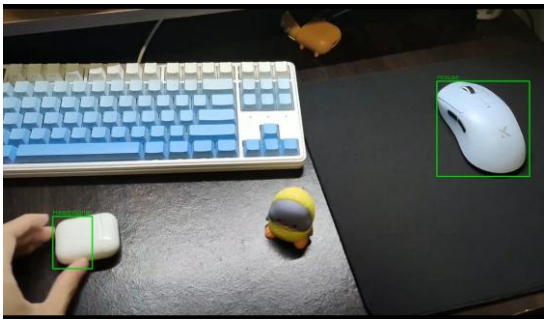


圖(一) 原始影片畫面

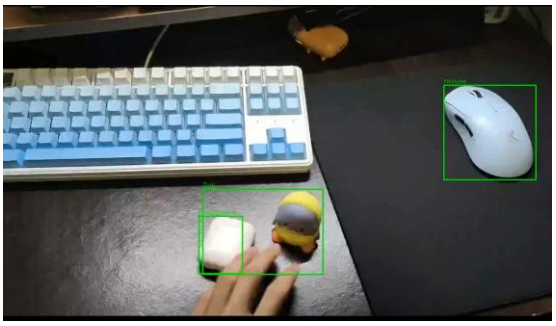
在 Github 中包含了不同 Prompt 的結果影片，在書面報告中我只抓取幾張照片進行呈現。可以從結果看到無論是輸入”Airpods”、”Headphone”，並且 Qwen 支援中文輸入，因此甚至是”耳機”、”白色物體”都能正確的抓取到 Airpods 這個物件。下表列出了不同提示詞(Prompt)的辨識結果，詳細的完整影片可以到 Github 中進行下載。



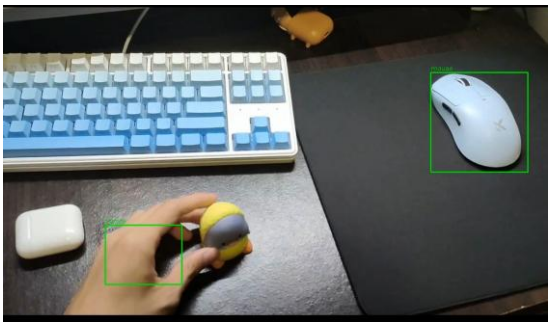
Prompt:Airpods、Mouse



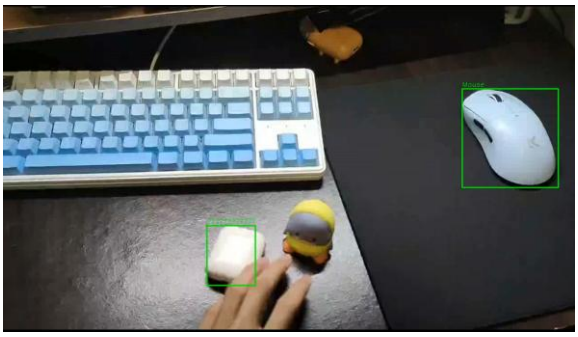
Prompt:Headphone、Mouse



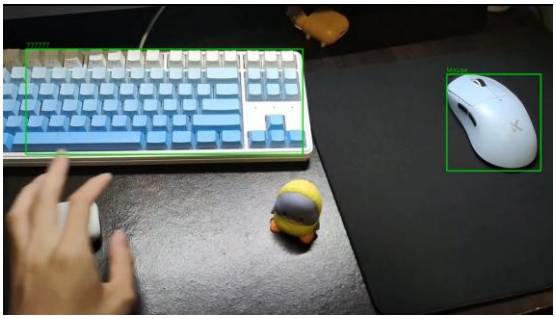
Prompt:Doll、Airpods、Mouse



Prompt:Hands、Mouse



Prompt:Mouse、白色物體



Prompt:鍵盤、Mouse

從上方的結果中可以看到，我的專案成功地達到了當初我們最初所期望的功能，透過 VLM 的幫助讓我們能夠在不需訓練模型的情況下，只需透過更改提示詞，就能讓系統追蹤不同的物件，我想已經取得了不錯的成效。但在實驗過程中會發現此種傳統物件追蹤的算法依舊有其侷限性，有時在相機移動過程中，系統會丟失追蹤的物件，或是在物體快速移動受到遮擋時，也會有追蹤錯誤的情況產生，因此在未來如果想要精進這套系統，或許可以朝著物件追蹤的穩定性進行改進。

四、 附錄：

專案資料:

1. Github Repo: https://github.com/Patrick-zhuyanxun/VLM-Multi-Object-Tracking.git

2. Google Colab 程式: https://colab.research.google.com/drive/1ykYKMwj9uwobXbj3zZJmLR9D7oMzQDf0?usp=sharing

參考資料:

3. Qwen2.5-VL:https://qwenlm.github.io/blog/qwen2.5-vl/
4. Localization code: https://github.com/QwenLM/Qwen2.5-VL/blob/main/cookbooks/spatial_understanding.ipynb
5. Object Tracking: https://medium.com/@khwabkalra1/object-tracking-2fe4127e58bf
6. CSRT: https://openaccess.thecvf.com/content_cvpr_2017/html/Lukezic_Discriminative_Correlation_Filter_CVPR_2017_paper.html