

多模態視覺 AI 與 Vision Agent 開發實戰

從 CLIP、SAM、VLM 到視覺智慧代理

■ 課程介紹

生成式 AI 正從文字生成快速邁向能同時理解文字、影像與影片的多模態人工智慧 (Multimodal AI)。從 CLIP 建立文字與影像的語意連結、SAM 進行提示式影像分割，到 Vision-Language Model (VLM) 進行影像問答與推理，AI 已從傳統影像辨識逐步發展為能「看懂、推理並執行任務」的視覺智慧系統。

隨著 AI Agent 發展，當 VLM 結合物件偵測、影像分割、生成式 AI 與工具調用，即可形成能理解視覺資訊並執行任務的 Vision Agent (視覺智慧代理)。

本課程從 Transformer、Vision Transformer (ViT) 出發，介紹 CLIP、Grounding DINO、SAM、BLIP、LLaVA、Qwen-VL 等 Vision Foundation Model 與 VLM 技術，並延伸至 Multimodal LLM、Tool Calling 與 Vision Agent。課程採「原理 + Python 實作 + 產業案例」，實作影像文字檢索、Zero-shot Classification、Open-Vocabulary Detection、Prompt-based Segmentation、VQA 與視覺推理，最終整合 VLM 與 Computer Vision Tools，建立可依影像內容選擇工具並完成分析任務的 Vision Agent。

課程並結合智慧製造、瑕疵檢測、智慧交通、自動駕駛及影像資料分析等應用，使學員從「影像理解」進一步掌握「視覺推理、工具調用與任務執行」的完整流程，建立新一代 Multimodal AI 與 Agentic AI 的實務應用能力。

■ 課程目標

1. 理解 CNN、Vision Transformer、Vision Foundation Model 與 VLM 的技術演進與核心概念。
2. 掌握 CLIP、Grounding DINO、SAM 等模型原理，以及 Zero-shot、影像文字檢索、物件偵測與影像分割等應用。
3. 理解 VLM 與 Multimodal LLM 架構，實作 Image Captioning、VQA 與 Visual Reasoning 等多模態視覺任務。
4. 了解生成式 AI 與 VLM 的整合方式，以及影像理解、生成與資料擴增等應用。
5. 掌握 AI Agent、Tool Calling 與 Vision Agent 核心概念，整合 VLM 與 Computer Vision Tools，建立簡易 Vision Agent。
6. 了解 Multimodal AI 與 Vision Agent 於智慧製造、自動駕駛及其他產業場景的應用。

■ 課程特色

1. 完整視覺 AI 技術鏈：串聯 CLIP、Grounding DINO、SAM、VLM 到 Vision Agent，系統化掌握新一代多模態視覺 AI 技術。
2. 開源模型實作導向：透過 Python / PyTorch 實作影像檢索、開放詞彙偵測、提示式分割、VQA 與視覺推理，強化實務開發能力。
3. 聚焦 Vision Agent 整合：整合 VLM、Tool Calling 與 Computer Vision Tools，實作視覺工具自主調用，延伸至智慧製造、視覺巡檢、智慧交通與自動駕駛等應用情境。

■ 課程對象

1. AI、Machine Learning、Deep Learning、Computer Vision 相關工程師與研發人員。
2. 從事影像處理、智慧製造、瑕疵檢測、自動駕駛、智慧交通及機器人等視覺 AI 應用之技術人員。
3. 希望從 CNN / YOLO 等傳統 Computer Vision 技術，進一步學習 CLIP、SAM、VLM 與 Multimodal AI 的工程師。
4. 希望將 VLM、Vision Agent 與生成式 AI 導入影像分析、智慧製造或自動化系統之工程師與技術主管。
5. 具備 Python 基礎，希望進一步學習多模態視覺 AI 與 Vision Agent 實作之學生、開發者或工程人員。

■ 課程大綱

主題	內容
從 Computer Vision 到 Multimodal AI	1. Computer Vision 發展趨勢 2. CNN 與 Vision Transformer 簡介 3. Foundation Model 概念 4. Multimodal AI 發展現況 5. CLIP 基本原理 6. Image-Text Matching 與 Zero-shot Learning
	實作：CLIP 影像理解 1. 使用 CLIP 進行影像分類 2. Image-Text Similarity 分析 3. Zero-shot Classification 實作
Vision Foundation Models	1. Open-Vocabulary Detection 概念 2. Grounding DINO 介紹 3. Segment Anything Model (SAM)介紹 4. Prompt-based Segmentation 5. Open-Vocabulary Vision Pipeline 建構
	實作：Grounding DINO + SAM 1. 文字指定任意物件 2. 自動物件偵測與分割 3. 建立 Open-Vocabulary 影像分析流程
Vision-Language Models(VLM)	1. VLM 核心架構與原理 2. BLIP、LLaVA、Qwen-VL 介紹 3. Visual Question Answering (VQA) 4. Image Captioning 5. Visual Reasoning 與 Scene Understanding 6. Multimodal LLM 發展趨勢
	實作：VLM 影像問答與推理 1. 圖片內容理解 2. 多輪影像問答 3. 工業影像與道路場景分析



主題	內容
VLM × Generative AI	1. Diffusion Model 基本概念 2. Stable Diffusion 與 DiT 簡介 3. Image Editing 技術介紹 4. Synthetic Data Generation 5. VLM 與生成式 AI 整合應用
	案例實作：Synthetic Data 1. 天候轉換案例 2. AI 資料擴增(Data Augmentation) 3. Synthetic Data 品質評估
Vision Agent 開發實戰	1. AI Agent 概念與架構 2. Tool Calling 與 Function Calling 3. Vision Agent 設計流程 4. VLM 與 Computer Vision 工具整合 5. 工具自主選擇、任務編排與 Agent 工作流程設計
	實作：建立 Vision Agent 1. VLM 理解任務需求 2. 自動選擇並呼叫偵測/割工具 3. 彙整工具結果 4. 自動產生分析結果與報告
Multimodal AI 產業應用與未來趨勢	1. 智慧製造與 AI 視覺巡檢案例 2. 智慧交通與自動駕駛案例 3. Vision Agent 產業導入情境 4. Vision-Language-Action (VLA) 5. World Model 與未來發展方向

★ 本課程實作以 PyTorch 深度學習框架為主，並於 Google Colab 雲端環境進行。

■ 講師簡介：

林哲聰 老師

學歷：台灣大學應用力學研究所碩士/清華大學資訊工程研究所博士

現職：瑞典自駕車軟體開發公司 Zenseact(Volvo Cars)深度學習工程師

經歷：馬來西亞偉特科技公司(ViTrox)研發顧問

馬來西亞 10 EPOCH 科技公司研發顧問

工研院機械所副研究員/研究員/資深研究員

加州大學聖塔芭芭拉分校資工系訪問研究員

上奇資訊-計算機概論 共同譯者

2010 伽利略創新大賽 台灣區季軍

2011 伽利略創新大賽 特別獎(GNSS Living Lab)得主

2013 伽利略創新大賽 瑞士區冠軍

2016 日本立命館大學英文演講比賽 清華大學代表

2017 CVGIP 行人偵測競賽 亞軍

2017 MOST 生成式對抗網路競賽 佳作

2018 第八屆兩岸清華研究生學術論壇 清華大學代表

2019 IEEE ICIP Three Minute Thesis Competition (3MT®): Finalist

2020 AI 智慧應用新世代人才培育計畫-人才解題實證(倒車攝影機影像識別-障礙物與逼近偵測)：佳作

2020 中華民國影像處理與圖形識別學會第十三屆博碩士論文獎博士論文佳作

專長：Computer Vision, Image Processing, Pattern Recognition, Machine Learning, Deep Learning, Python/C/C++ Programming

已於電腦視覺/深度學習/駕駛輔助/生成模型/自駕車領域中發表過 37 篇國際論文，15 篇國內論文，以及 13 篇專利



【開課資訊】

- **主辦單位：**工研院產業學院 台北學習中心
- **舉辦地點：**實體：工研院產業學院 產業人才訓練一部(台北) (實際地點以上課通知為主)
線上：WEBEX 線上會議室 (將於課前寄發信件通知)
- **上課時間：**116/2/18、19 (四、五)，09:30~16:30 (共 12 小時)
- **招生人數：**本班預計 20 人為原則，最低開課門檻為 10 人
- **課程費用：**

報名方案	費用
課程原價	每人 12,000 元
早鳥價 (開課前 3 週)	每人 10,800 元
團報價 (三人以上)	每人 10,200 元

- **報名方式：**線上報名：點選課程頁面上方的「線上報名」
- **繳費方式：**報名時可選擇信用卡線上繳費或匯款，主辦單位將於確認開班後通知您相關匯款資訊
- **退費標準：**若欲取消報名，請於開課前三日以 email 告知主辦單位，並電話確認退費事宜，逾期將郵寄講義，恕不退費。若您未於期限內申請退費，則不得於任何因素要求退費，惟可轉讓與其他人參訓
- **課程洽詢：**02-2370-1111 *357 凌小姐
- **注意事項：**
 1. 為確保您的上課權益，報名後若未收到任何回覆，敬請來電洽詢方完成報名。
 2. 若原報名者因故不克參加，但欲更換他人參加，敬請於開課前三日通知。
 3. 因課前教材、講義及餐點之準備，若您不克前來需取消報名，請於開課前三日以 EMAIL 通知主辦單位聯絡人並電話確認申請退費事宜，逾期將郵寄講義，恕不退費。
 4. 配合講師時間或臨時突發事件，主辦單位有調整日期或更換講師之權利。
 5. 報名時請註明欲開立發票完整抬頭，以利開立收據；**未註明者，一律開立個人抬頭**，恕不接受更換發票之要求。
 6. 為尊重講師之智慧財產權，課程進行中請勿錄音及錄影。