結合大語言模型與視覺特徵之
多模態食物語義分割系統
LLM-Assisted Multimodal Semantic Segmentation for Food Image Analysis

指導教授：朱威達

專題成員：戚瑞豐、林聖隆

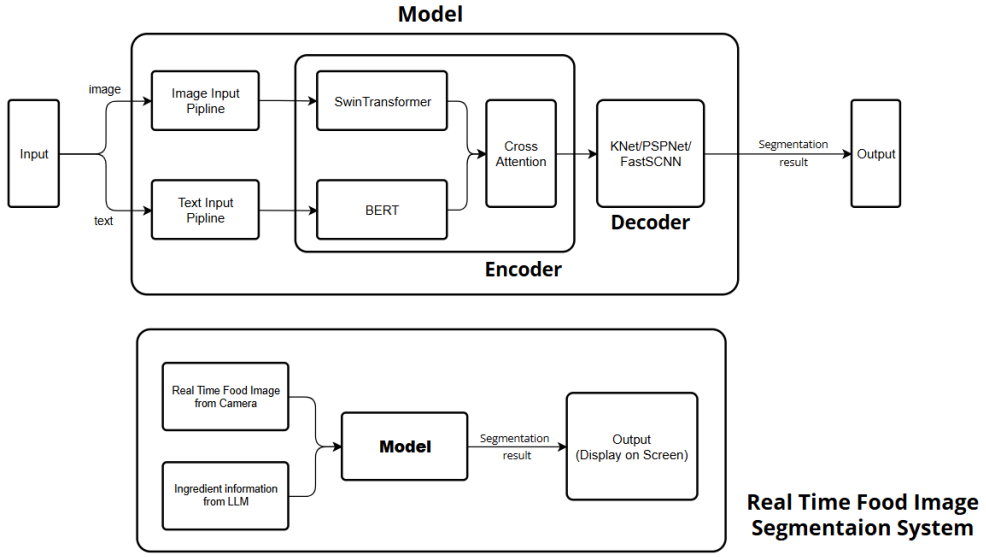
開發工具：MMSegmentation, FoodSeg103, Pytorch

測試環境：CUDA11.3, mmsegmentation1.1.2

一、簡介：

本專題旨在開發一結合大語言模型（LLM）與視覺特徵的多模態食物語義分割系統。隨著健康意識提高，實時食物分析的需求日益增長，然而傳統的食物圖像分割技術主要依賴視覺特徵，對於外觀相似但屬於不同類別的食物常出現誤判。為解決此問題，我們在原有的 Real-Time Food Image Segmentation 系統基礎上，引入大語言模型產生的標籤訊息（label messages），將文字特徵與圖片特徵進行交叉融合（cross-feature），從而提升食物分割的準確性。本系統採用 FoodSeg103 資料集進行訓練，使用 MMSegmentation 開發平台實現三種分割模型（K-Net、PSPNet 及 Fast-SCNN），經比較選出最佳模型後整合至相機系統，實現實時食物分析功能。

以下為系統架構圖（圖一）：



（圖一）

專題的系統架構圖 – 自行繪製

二、測試結果：

2.1 模型評估結果：

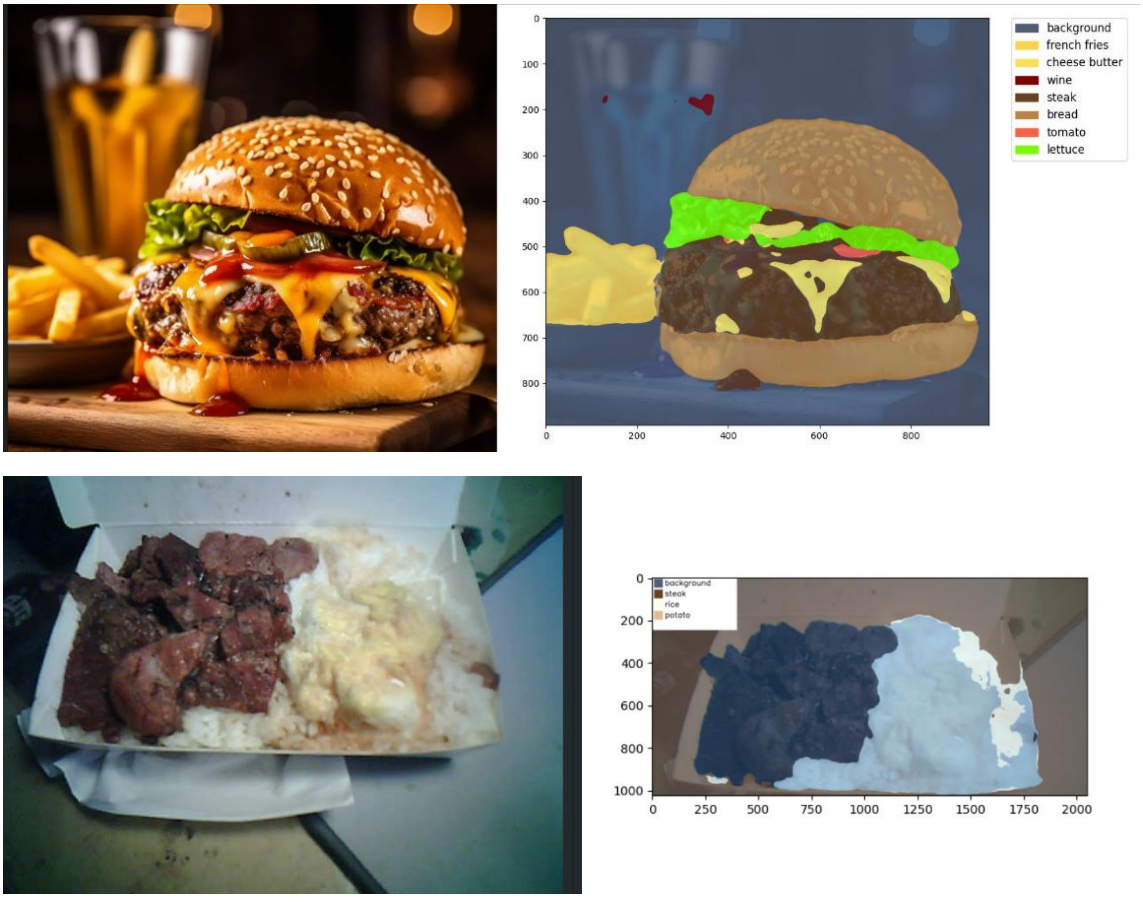
本研究使用主要使用「平均交叉並集（mIoU）」來衡量模型性能。下表顯示了三種基礎模型與加入 LLM 產生標籤訊息後的性能比較：

模型	原始mIou	LLM增強後 (依現場展示為準)	提升幅度
K-Net	37.01	約 41.4	**+11.91%**
PSPNet	15.19	約 17.0	**+11.98%**
Fast-SCNN	2.84	約 3.2	**+13.73%**

結果顯示，結合大語言模型產生的文字特徵後，所有模型的分割性能都有明顯提升。其中 K-Net 模型整體表現最佳，並被選為最終的部署模型。

2.2 食物分割效果展示

下圖（圖二）為系統在對於「靜態與動態圖片」中的食物分割效果展示。我們透過 K-Net 的模型進行語義分割：



（圖二）

靜態與動態圖片的食物分割結果 - 由系統所產生