

2025年臺灣國際科學展覽會 優勝作品專輯

作品編號	190028
參展科別	電腦科學與資訊工程
作品名稱	漫畫生成與預測

就讀學校	國立臺灣師範大學附屬高級中學
指導教師	李柏翰 李啟龍
作者姓名	蔡昕翰 葉安之 吳宇翔

關鍵詞	<u>生成式人工智慧、Low Rank Adaptation</u>
-----	------------------------------------

作者簡介



我們是來自師大附中資訊科學班的高三同學，對於資訊科技充滿熱忱，我們專攻於後端開發及人工智慧模型訓練。在這次國際科展中，我們將創意與技術結合，提出了一個能生成漫畫劇情的解決方案，名為「漫畫生成與預測」。該研究以生成式人工智慧技術為基礎，設計出三步驟的生成流程，包含劇情提取、預測與圖像生成，為漫畫創作提供全新可能性。

附件三

研究報告封面

2025 年臺灣國際科學展覽會

研究報告

區別：

科別：電腦科學與資訊工程科

作品名稱：漫畫生成與預測

關鍵詞：生成式人工智慧、Low-Rank Adaptation

編號：

（編號由國立臺灣科學教育館統一填列）

摘要

本研究探討了利用生成式人工智慧技術為漫畫創作帶來新可能性。在當今競爭激烈的漫畫產業中，創作者們需要不斷創新以吸引觀眾，而創作引人入勝的漫畫需要豐富的想像力和劇情結構。本研究希望能協助創作者製作草稿，並探索與 AI 當朋友的新型創作模式。在生成方面，提出了將漫畫劇情提取、劇情預測以及圖片生成三個步驟的生成流程，並使用了多種模型和技術，如 YOLO 模型用於漫畫人臉檢測、文字生成模型用於劇情預測、LoRA 技術用於模型微調等，為解決人物生成不連續的問題，我們也提出一種基於特徵提取與融合的解決辦法。本研究提供了一個全面的方案，旨在利用人工智慧技術幫助漫畫創作者創作出簡單的草稿。

Abstract

This study explores new possibilities for comic creation using generative AI technology. In today's highly competitive comic industry, creators must constantly innovate to captivate their audiences, as creating engaging comics requires a rich imagination and well-structured plots. The goal of this research is to assist creators in drafting and to explore a new collaborative creation model where AI serves as a partner. The proposed generation process involves three key steps: comic plot extraction, plot prediction, and image generation. Various models and techniques are employed, such as the YOLO model for comic face detection, text generation models for plot prediction, and LoRA technology for model fine-tuning. To address the issue of character generation continuity, we also propose a solution based on feature extraction and its combination. This study provides a comprehensive framework designed to help comic creators generate simple drafts using AI technology.

壹、 前言

一、 研究動機

在當今日益蓬勃的漫畫產業中，創作者們不斷尋求創新和突破，以吸引觀眾並提供令人驚艷的劇情。然而，創作一個引人入勝的漫畫並非易事，需要豐富的想像力、情節結構和角色塑造。在這個背景下，開發一個可預測漫畫的生成式人工智慧，幫忙打草稿。

二、 研究目的

- (一) 協助創作者製作草稿，為作者減少靈感枯竭的痛苦
- (二) 探索新型創作模式，讓一般人可以更接近漫畫創作
- (三) 讓中途停刊的漫畫得到一線生機

貳、 研究方法

一、 系統流程

我們的生成流程如圖。

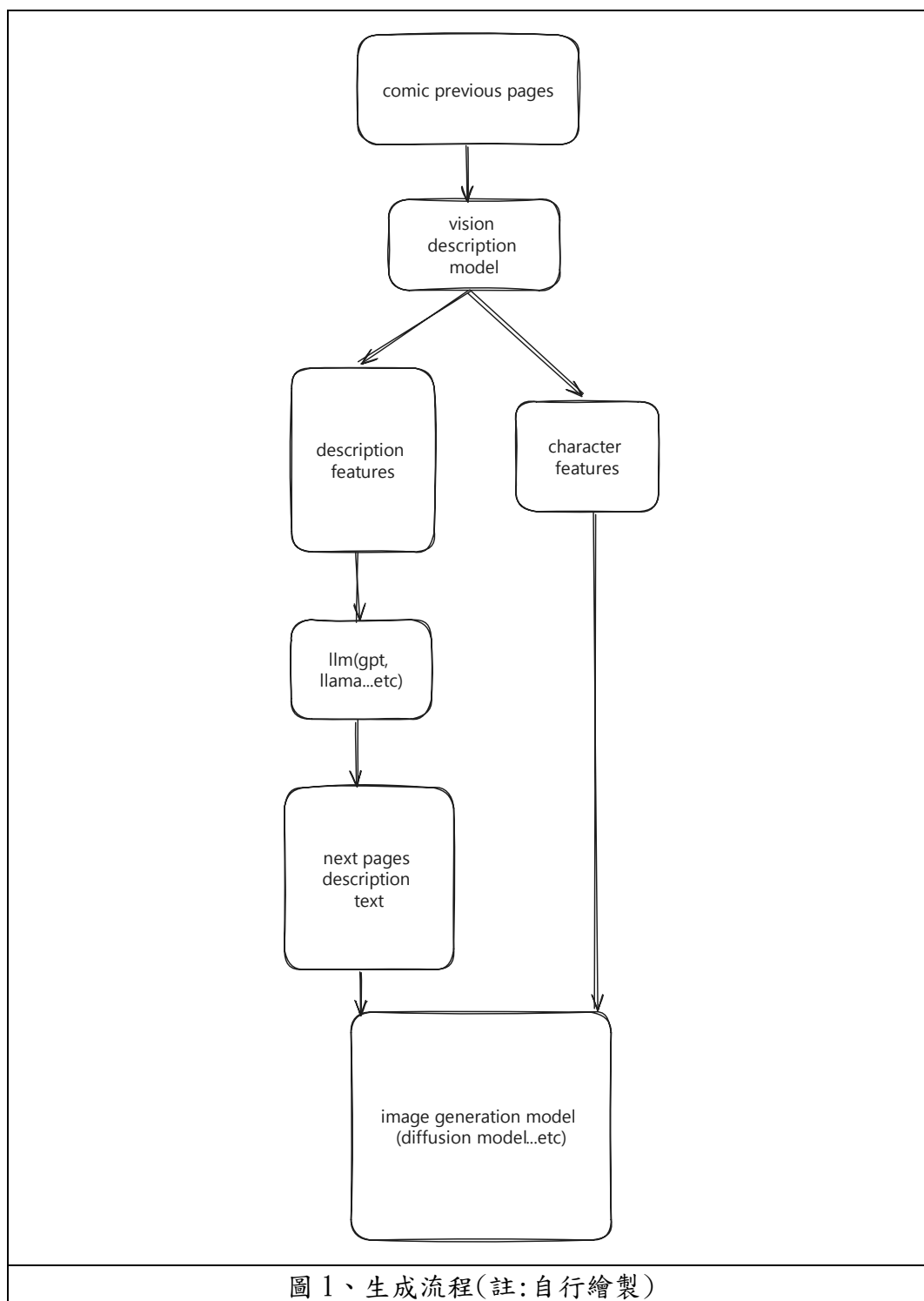
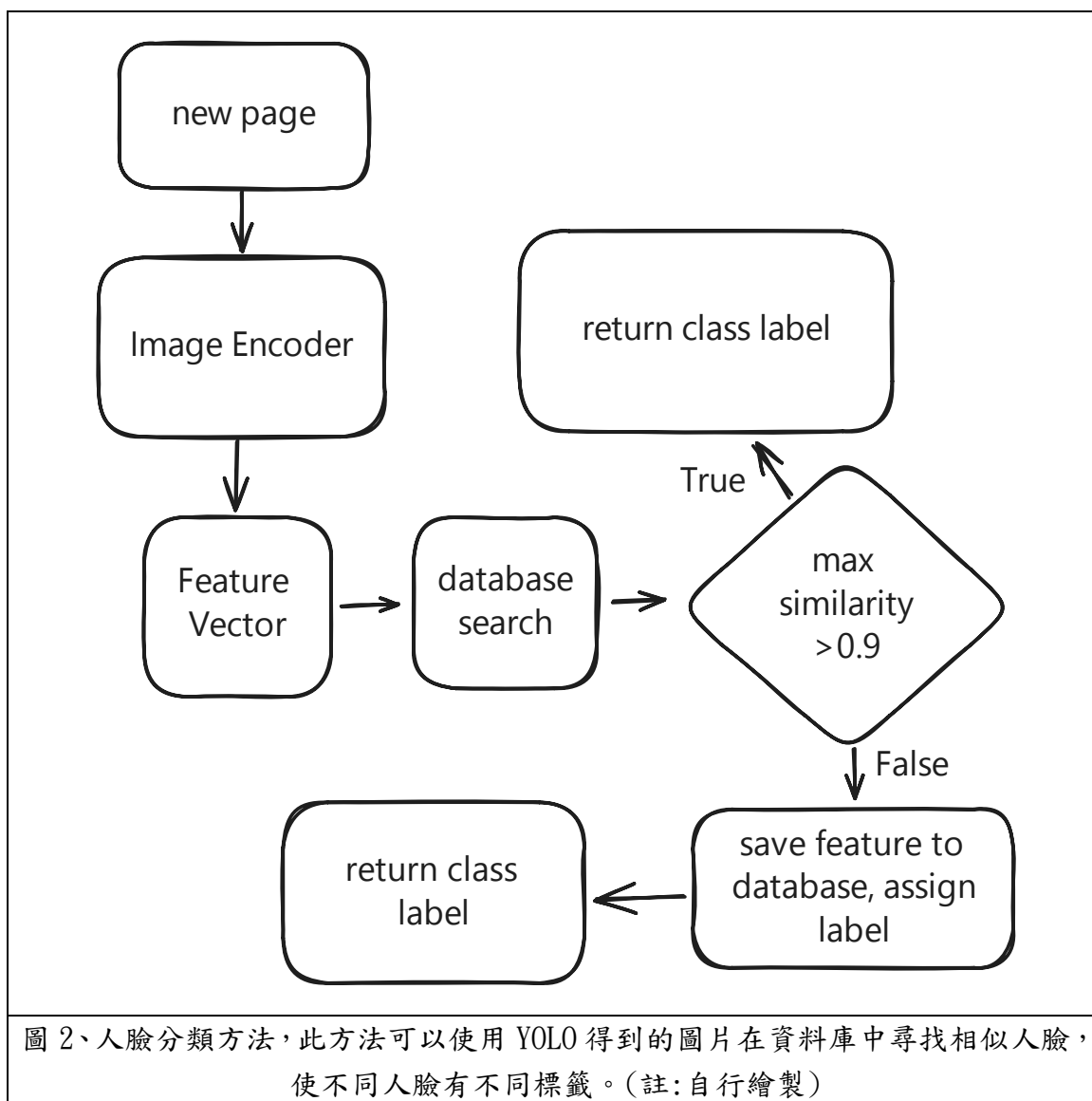


圖 1、生成流程(註:自行繪製)

這個流程圖可以把我們最大的問題——也就是將一串漫畫生成出下一張漫畫的圖片——分成三個較小的問題。第一個小問題，我們會使用的是影像敘述 AI，讓模型把各張漫畫輸出敘述性文字，在這個步驟，我們可能會用到不只一種模型，例如「某某說了『我要統治世界』這句話」，這個我們會用 YOLO 先將漫畫中的人臉辨識出來[1]，用孿生神經網路判斷是哪一個角色，再敘述每個角色的外觀，並用 ocr 將文字也擷取出來，將其位置映射到最近的人臉。[2]



當然這個方法會有問題，因為有時候對話框會互相交叉，造成離其他角色更近，有時候也會把場景/劇情敘述當成對話框，因此我們找到了一種方法，先用 YOLO 把文字框找到，再用 OCR 辨識框內文字。

第二步驟是使用大型語言模型將漫畫的劇情預測出下一張的劇情，這個部分沒有太大的問題，我們也就只有一些選擇像是 LLaMA 系列、GPT 或 Grok，就可以完成這個任務，我們再自己訓練一個比他們好的模型也機會渺茫，因此第二步驟是最簡單的一步。[3]

但第三步驟是將敘述性文字輸出圖片，這是最困難的一個步驟，因為沒有很多人有做過，所以沒有開放的模型給我們使用。這個模型的輸入我們會需要敘述文字和涉及到的人物的圖片。根據我們目前查到的資料，我們可以使用 character features(角色特徵)與 LoRa 方法，遷移學習 stable diffusion 或 Midjourney 使其能生成出人物外觀與漫畫人物較相似的漫畫。[4]

二、細部架構

(一) YOLO 模型

YOLO (You Only Look Once) 是一種流行的物體檢測模型，其特點是高速和高準確率。相較於傳統的物體檢測方法，YOLO 模型在一次前向運算中直接預測多個邊界框和類別概率，因此被稱為 "you only look once"。YOLO 模型通常基於卷積神經網路 (CNN) 構建。它採用單一的神經網絡，將輸入圖像分割成網格，每個網格預測多個邊界框和對應的物體類別機率。與其他方法不同，YOLO 僅通過一次前向傳播運算，即可直接預測出圖像中所有物體的位置和類別。YOLO 模型利用多層次的特徵來預測物體的位置和類別。這使得它能夠有效地檢測各種尺寸和形狀的物體。其使用綜合的損失函數來訓練模型，包括位置損失和分類損失。這使得模型能夠在訓練過程中同時優化物體的位置和類別預測。[1]

將 YOLO 模型應用於漫畫中人臉的檢測通常需要這些步驟：

首先，需要使用漫畫圖像數據集對 YOLO 模型進行訓練。訓練集應包含標記有人臉位置的圖像，並對模型進行監督學習。訓練完成後，將模型應用於漫畫圖像，通過 YOLO 模型預測出圖像中的人臉位置。在檢測到的人臉位置上進行後處理，例如過濾掉重疊的框、根據模型的信心分數 (confidence score) 選擇最正確的檢測結果等。最後，可以根據檢測到的人臉位置進行相應的應用，例如人臉識別、情感分析等。

在漫畫中檢測人臉時，由於漫畫風格的多樣性和複雜性，可能需要對 YOLO 模型進行一定程度的調整和優化，以提高檢測的準確性和穩健性。[5]



圖 3、我們使用 YOLO 辨識出漫畫人物位置(註:自行繪製+One Piece 宣傳海報)

(二) 文字生成模型

文字生成模型是能夠根據給定的上下文訊息生成自然語言文句的人工智慧模型。這些模型通常基於深度學習技術，例如遞歸神經網絡 (RNN)、長短期記憶網絡 (LSTM)、Transformer 等。

這些模型的作用是根據給定的輸入訊息 (例如一段文字、圖像、聲音等) 生成一段自然語言文本。它們可以應用於多種任務，包括機器翻譯、對話生成、情感分析、摘要生成等。在漫畫劇情預測中，文字生成模型的作用是根據前幾張漫畫的內容生成下一張漫畫的劇情描述。

文字生成模型可以根據前幾幅漫畫的內容，生成下一幅漫畫的劇情描述。這些描述可以涵蓋角色的行為、對話、情緒等訊息，有助於推動劇情的發展。他也可以確保生成的劇情描述與前幾幅漫畫的內容相關聯，從而保持劇情的連貫性。這有助於讓劇情更加流暢、自然。根據前幾幅漫畫的內容也可以生成多種可能的劇情描述，為創作者提供靈感和想法。這有助於創作者發掘新的劇情發展方向，使作品更加豐富和有趣。[6]

(三) LoRA 模型

針對巨大的語言模型，研究人員開發了一種新穎的技術，名為低秩適應 (LoRA)，可以高效地對這些模型進行微調。傳統的微調方法需要更新模型中所有的參數，計算量龐大。LoRA 的做法是只更新一小部分稱為 "適配器" 的低秩矩陣，而保持原始模型權重不變，如下圖 4 可見：

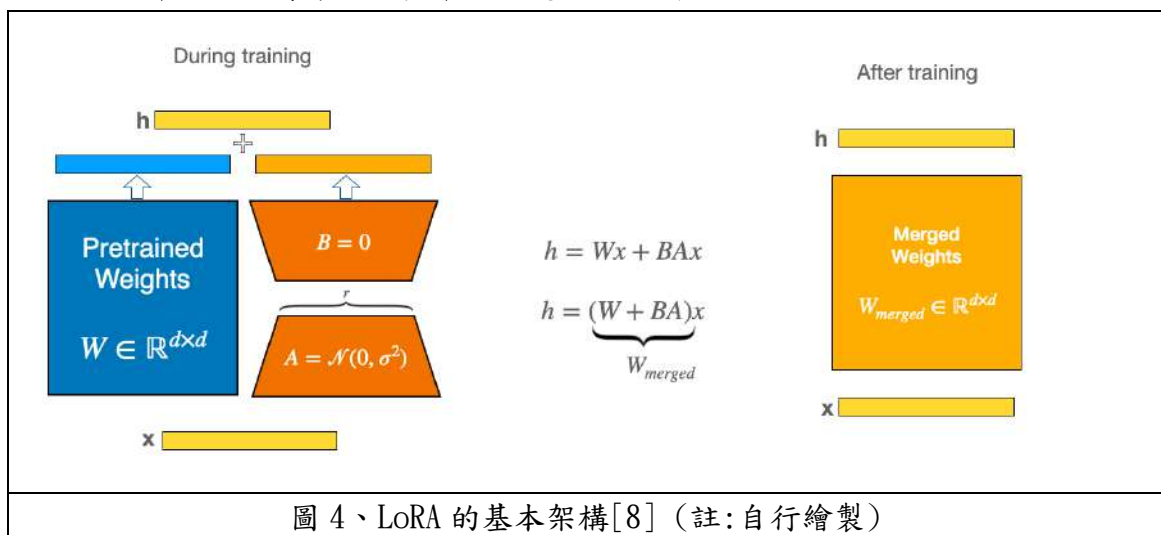


圖 4、LoRA 的基本架構[8] (註: 自行繪製)

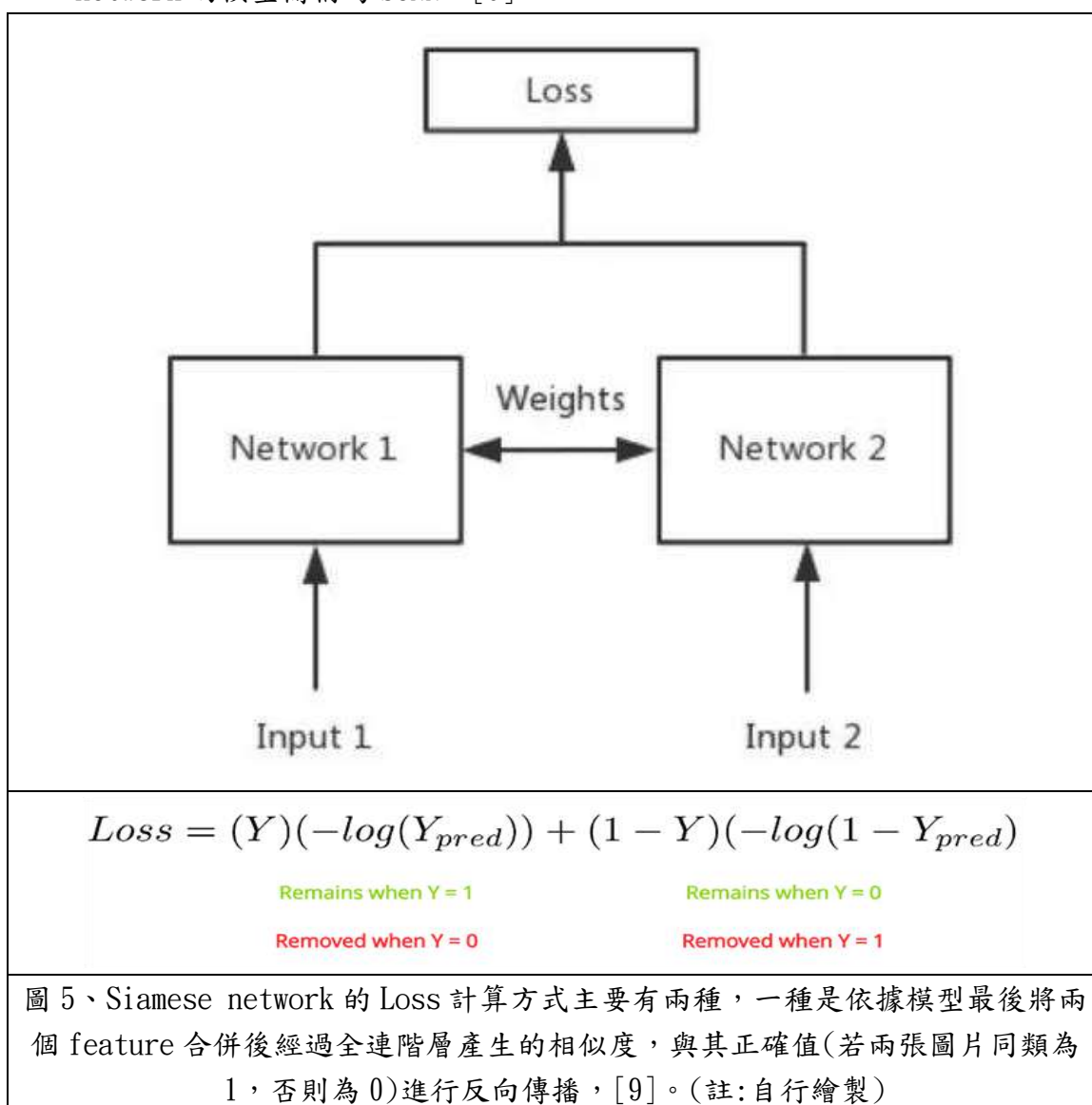
這些適配器矩陣的參數量雖然比要訓練的模型本身少很多，但卻能充分利用預訓練模型中豐富的知識，使微調後的模型在新任務上表現出色，接近完整微調的水準。由於只需調整少量參數，LoRA 的計算和內存消耗遠低於傳統微調，能夠快速高效率地將通用語言模型專門化到特定領域。[7, 8]

LoRA 雖然一開始是用於語言模型的微調方法，不過他也可以推廣到其他大參數量模型。我們預計使用 LoRA 搭配每個漫畫角色的特徵與資料集訓練

一個生成以外觀相似的人為主角之漫畫的模型，像是 stable diffusion 或 Midjourney，因為我們看到相似研究使用相同方式有不錯結果。

(四) 孿生神經網路

我們有使用一種叫做 Siamese network 的模型，此模型簡單來講就是有兩個相同權重的 CNN(Convolutional Neural Network，卷積神經網路，此處泛指任何使用卷積層進行影像分類的模型)，將兩張圖片進行特徵提取，並對其進行比對相似度以達到僅需少樣本甚至單樣本辨識。因為這兩個模型的權重是一樣的，而有了 Siamese(孿生)的名字。這種使用 CNN 進行 Siamese network 的模型簡稱為 SCNN。[9]



另一種是使用 Triplet loss，挑選與輸入照片同類、不同類的資料各一個，計算出原照片與他們的歐式距離並透過 loss function 促使其較接近同類的照片，[10]。

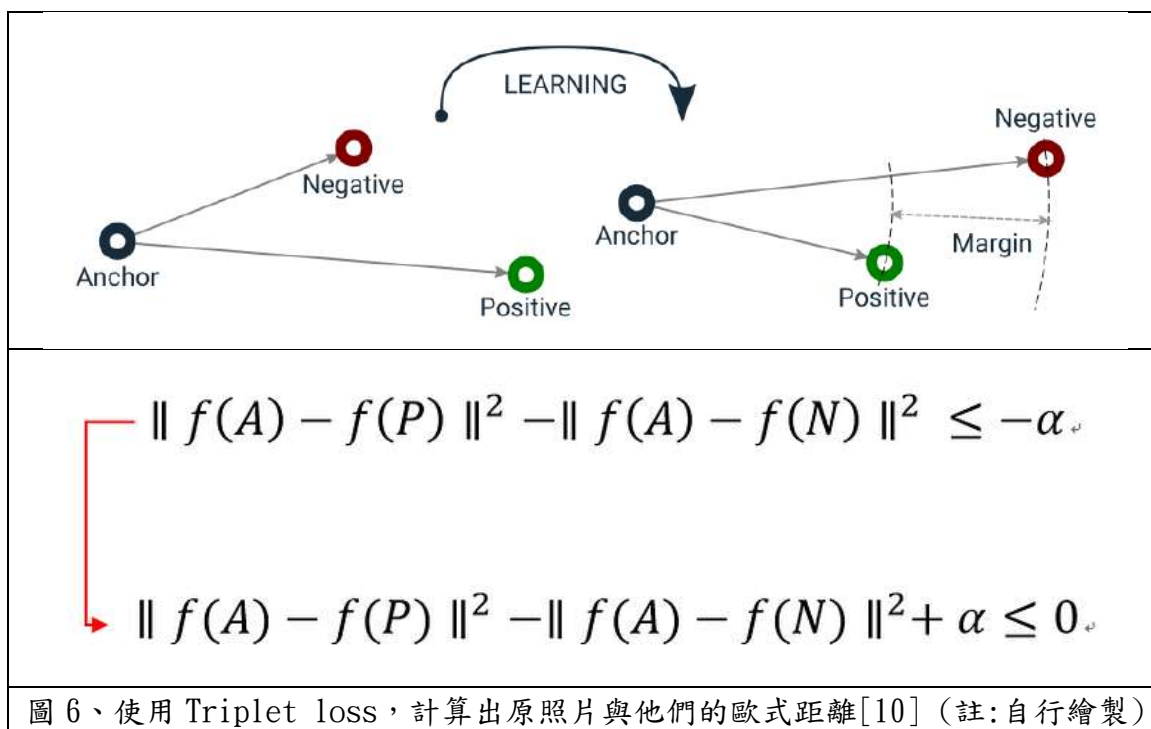


圖 6、使用 Triplet loss，計算出原照片與他們的歐式距離[10]（註：自行繪製）

(五) 對話框文字擷取

在計畫初期，我們使用的是 Tesseract OCR，不過這個版本的 OCR 在辨識我們的對話框時表現不佳（辨識度低、辨識錯地方等問題），因此我們找到另一個方法，先用 YOLO 辨識出對話框位置，再依語言將對話框內的文字題取出來。[11]

(六) 以漫畫圖中特徵向量生成人物有連續性的圖片

因為漫畫常常會出現新的角色，如果還要去漫畫中搜尋每個角色的照片拿來訓練 LoRA，手動很麻煩，但自動化又很容易出錯，因此我們想到可以開發一個模型能將角色圖片轉換成特徵向量，而另一個模型可以直接用角色特徵向量以及描述角色姿勢、行為、背景的文字來生成圖片，此方法能不再重複訓練其他模型，節省運算資源與空間。

1. 我們首先會用 GAN 的方法訓練兩個模型，透過「特徵化」再「圖像化」出的圖片與原圖對比相似度來訓練「特徵化」與「圖像化」的模型，使用的圖片是去掉背景且只包含單一角色
2. 訓練好「特徵化」與「圖像化」的模型，就可以來訓練 diffusion model 根據角色特徵向量以及描述角色姿勢、行為、背景的文字生成圖片，希望可以成功修改架構、訓練模型接受多個角色特徵向量，繪出多人的圖片。

還有另一個想法是疊合圖片，因為我們在網路上查到有 pose-gan 等模型能夠以角色圖片生成出角色在不同姿勢的模樣，因此我們覺得有另一方式：可以訓練一個語言模型將輸入文字分解成各個角色的姿勢及行為用 pose-gan

等模型生成個別角色圖片、用另一個影像生成模型生成背景以及其他物品、最後再訓練模型把圖以正確位置、方向、順序疊合。[12]

參、 初步研究結果

首先，我們把漫畫中的文字與人物特徵提取出來，並寫成對話的形式，如表一。

表一、輸入資料	
第一頁	
輸入的圖片(註:圖片來源為 One Piece 試閱)	 The image is a page from the One Piece manga. It shows a scene with several characters and a large mechanical structure. The characters are speaking, and the structure is labeled 'IRON GIANT'. The text in the image is as follows: - ABSOLUTELY!! I'M VEGAPUNK TOO, YOU KNOW!! - YOU'RE CERTAIN? - I KNOW WHERE THE TRANSPONDER SNAIL IS!!! - FIVE ELDERS !! - THE TRANSMISSION TRANSPONDER SNAIL... ...OF THE IRON GIANT IN THE FABRIO-PHASE!!! ...IS UNDER THE PROTECTION... ?!! - NOTHING'S WORKING ON IT!! - BOOM!!
Ocr +llama 潤飾+漫畫人臉描述的輸出	old man 1: absolutely!! I' m vegapunk too, you know!! old man 1: the transmission transponder snail something: you' re certain? old man 2: I know where the transponder snail is!! a woman: five elders!! giant armor: ... of the iron gian in the fabrio phase!!! giant armor: ... is under the protection

	giant armor: nothing working on it!!
第二頁	
輸入的圖片(註:圖片來源為 One Piece 試閱)	
Ocr +llama 潤飾+漫畫人臉描述的輸出	from the building:Hey, Ray!! from the building:Why didn' t they act on it? from the building:...Why did they drift apart without telling anyone? from the building:when they learned all of the true history... old man: and you' re telling them too much vegapunk old man: don' t be a fool! us old folds should keep our secrets. old man: I' m just an old man now... a lady: It' s not like you to be this wiped out to have so much drink.

	old man: don' t rob the young' uns of all the fun. Right, roger?
--	--

接著我們用 LLaMA 3 由這幾頁的敘述生成出下一頁的結果，如表二。

表二、由表一的敘述輸出產生的下一頁漫畫敘述	
Page 3 Description: The scene shifts to a bustling laboratory filled with high-tech gadgets, glowing monitors, and bubbling test tubes. Several young scientists and engineers, clad in lab coats, are seen working diligently at their stations. In the center, a large holographic display shows a complex blueprint of a transponder snail. The atmosphere is tense but focused. Dialogue: **Young Scientist 1:** *pointing at the blueprint* "We've identified the malfunction in the transponder snail's circuitry." **Young Scientist 2:** "Excellent! We need to fix it before the next transmission." **Old Man (Vegapunk):** *entering the lab, looking more serious* "Remember, the secrets this snail holds are beyond comprehension. Handle it with the utmost care." **Young Scientist 3:** "Don't worry, Vegapunk! We've got this under control." **Old Man (Vegapunk):** *sighs* "I hope you're right. The fate of many depends on this." **Lady Scientist (earlier woman):** *approaching Vegapunk with concern* "Are you sure you're okay? You seemed quite distressed earlier." **Old Man (Vegapunk):** *forcing a smile* "Just reminiscing about old times. Let's focus on the task at hand." **From a Monitor (transmission):** "Attention! Unauthorized access detected in the east wing!" **Young Scientist 1:** "Security breach! Everyone, stay alert!" **Old Man (Vegapunk):** "It's starting. Prepare for anything!" The scene ends with the lab personnel scrambling to secure the area, while the old man watches, deep in thought, contemplating the gravity of the situation.	

表三、我們運用 LoRA v.s. 沒有 LoRA 的輸出效果差異	
Prompt	Luffy, portrait, (masterpiece, good quality)

無 LoRA 的輸出
(註:自行使用模
型繪製)



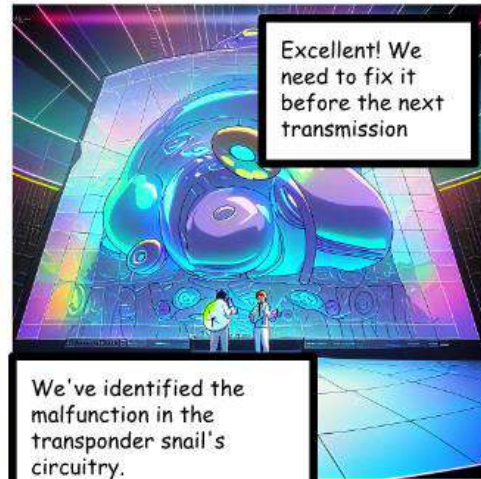
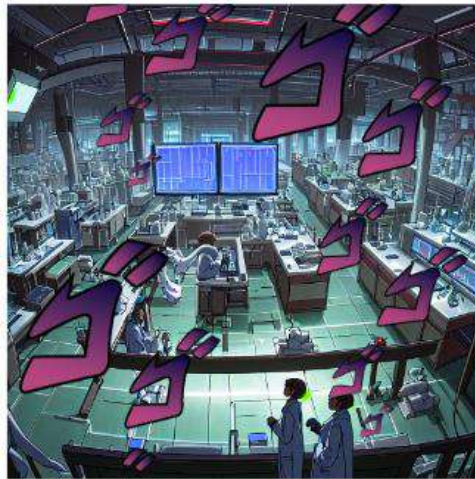
有 LoRA 的輸出
(註:自行使用模
型繪製)



最後我們用 LoRA 微調過後的 StableDiffusion XL 生成出六格漫畫。

表四、由表二的文字輸出敘述轉換為漫畫的圖

漫畫生成結果
(註:自行使用
模型繪製)



我們還有用 Yolo+OCR 的方法可以解決直接辨識辨識度不足的問題，先用 YOLO 將漫畫對話框擷取出來，再將框內文字辨識，效果如圖7。

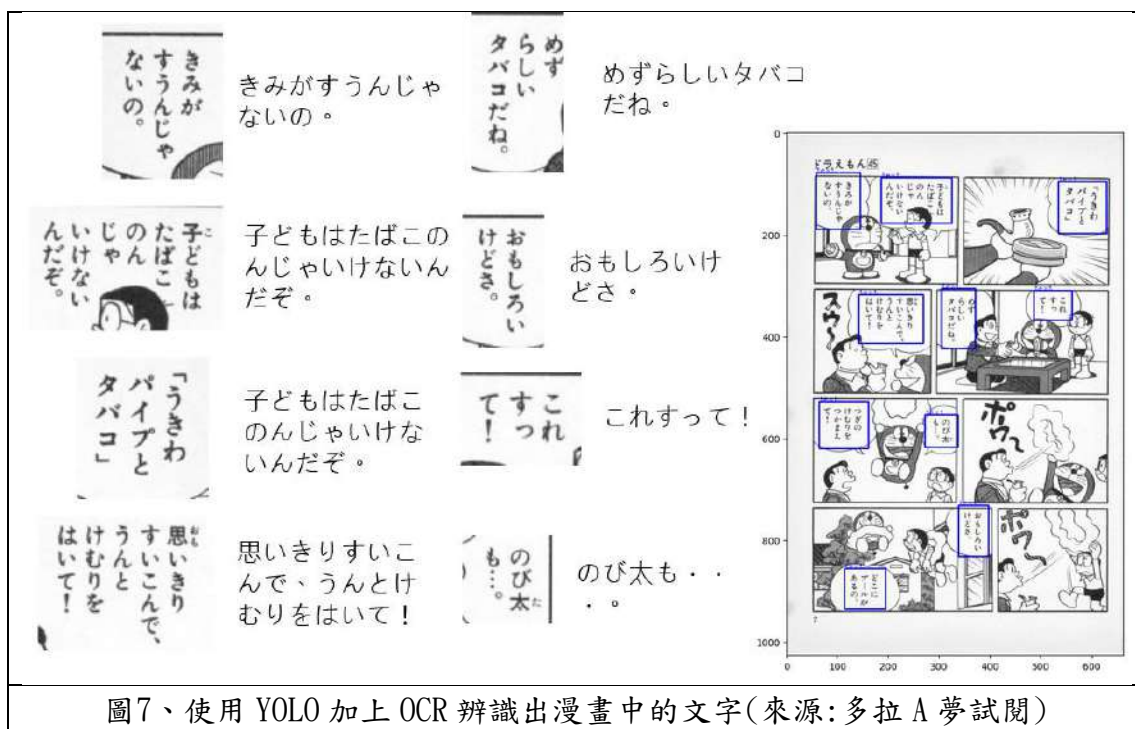


圖7、使用 YOLO 加上 OCR 辨識出漫畫中的文字(來源:多拉 A 夢試閱)

肆、 討論

使用 Tesseract OCR+基礎影像前處理在處理對話框時，經常會將兩個相鄰的對話框混淆在一起，這使得文句的理解和分割變得更加困難。此外，這個 OCR 在辨識漫畫字型方面表現出較低的準確度。使用 YOLO 先將文字框位置辨識出來，再用 EasyOCR 替代 Tesseract OCR 能解決上述問題。除此之外，人物的特徵擷取的問題也值得探討，使用這個系統生成的圖片中角色之間缺乏連續性和一致性。這種缺乏連貫性的問題會影響故事的流暢性和讀者的理解，尤其是在處理情節複雜的漫畫時，這一問題尤為突出。為此，我們提出了一個以提取漫畫人物特徵後結合文字生成圖片的解法，期以此方法得到更好的結果。

伍、 結論

我們成功開發出一個方法使其能擷取一序列漫畫中的特徵，結合並分析後產生出相對應之下一格合理之劇情，對於漫畫人物不連續的問題，也提出一個方法希望能解決，本研究能讓漫畫界的效率提升，產生出更多有趣的作品。

陸、 參考文獻

- [1] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).
- [2] Largecats. (n.d.). Largecats/comics-OCR: Tool for extracting script from comic pages using OCR engine tesseract. GitHub.

<https://github.com/largecats/comics-ocr>

Tesseract-Ocr. (n.d.). Tesseract-OCR/tesseract: Tesseract open source OCR engine (main repository). GitHub.

<https://github.com/tesseract-ocr/tesseract>

[3] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... Scialom, T. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv [Cs.CL]. Retrieved from <http://arxiv.org/abs/2307.09288>

[4] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10684–10695).

[5] Justsecret123. (n.d.). Justsecret123/one-piece-characters-detector: An object detector trained with a kaggle GPU, using tensorflow and a fine-tuned SSD RESNET50. GitHub.

<https://github.com/Justsecret123/One-Piece-characters-detector>

[6] Jin, Z., & Song, Z. (2023). Generating coherent comic with rich story using ChatGPT and Stable Diffusion. arXiv [Cs.CV]. Retrieved from <http://arxiv.org/abs/2305.11067>

[7] Tuning.AI. (2023, September 14). Lora: Low-rank adaptation - efficient fine-tuning for large language models. LinkedIn.

<https://www.linkedin.com/pulse/lora-low-rank-adaptation-efficient-fine-tuning-large-language>

[8] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. arXiv [Cs.CL]. Retrieved from <http://arxiv.org/abs/2106.09685>

[9] Koch, G., Zemel, R., & Salakhutdinov, R., Siamese neural networks for one-shot image recognition. In ICML deep learning workshop Vol. 2(1), (2015).

[10] Hermans, A., Beyer, L., & Leibe, B. In defense of the triplet loss for person re-identification. (2017) arXiv preprint arXiv:1703.07737.

[11] ogkalu2, Comic Translate, (2024), GitHub repository,

<https://github.com/ogkalu2/comic-translate>

[12] AliaksandrSiarohin. Deformable GANs for Pose-based Human Image Generation. (2017) Github repository,

<https://github.com/AliaksandrSiarohin/pose-gan>

【評語】190028

本作品具有創意，運用微調技術與移轉學習的深度學習模型，讓 A I 生成六格漫話，並能從漫畫對話文字中擷取出文字。

建議可以增加人為評估實驗，以了解生成器所產生的漫畫品質與人為漫畫間的差異作為比較與回饋精進。