國立清華大學 電機工程學系

實作專題研究成果報告摘要

Real-Time Chatroom Sticker Generation with Image Generative AI

影像生成式 AI 於聊天室貼圖即時生成應用

專題領域：資工領域

組　　別：B463

指導教授：黃朝宗

組員姓名：林庭伃、徐瑞澤

研究期間：113 年 1 月 20 日至 113 年 11 月 20 日止，共 10 個月

摘要

日常生活中，許多人傳訊息時會使用「貼圖」輔助表達，但有時候卻面臨找不到合適貼圖的困境；且即使貼圖創作者時常更新，單一系列貼圖也難以涵蓋所有使用者想傳達的意涵，使「找不到貼圖」的困境始終存在。面對這個問題，圖像生成式模型是我們所想到的解方。

近年來圖像生成式模型發展迅速，而擴散模型(Diffusion Model)因穩定的優良表現成為關注焦點；OpenAI 的 DALL-E、Google 的 Imagen 和 StabilityAI 的 Stable Diffusion 等模型皆採用此架構，並獲得相當好的迴響。然而，因訓練生成式模型要耗費大量的運算資源及時間，因此原先用於大型語言模型(Large Language Model, LLM) 訓練的 Low-Rank Adaptation (LoRA)，被引入到擴散模型訓練，以降低運算需求。

我們使用多位創作者製作的 LINE 貼圖作為訓練的 Dataset，並採取 LoRA 訓練的方式，用相對較短的時間與較少的硬體使用量，針對不同的創作風格訓練模型，以此生成各式各樣符合表情、語句的貼圖。此外，因我們想達成的目標為即時貼圖產生，故選擇 inference time 較短的開源模型 SDXL-Turbo 作為訓練的對象。

在此專題研究我們成功產生風格明確且表情豐富的貼圖，並且展示在相同提示詞(prompt)下生成貼圖的多樣性。除了訓練不同風格的模型外，我們也利用 Fuse LoRA，將不同模型的風格融合，以此迸發出新風格。

為了展示我們模型於生活中的應用，我們結合 Huggingface Diffusers、Transformers 函式庫與 Web Framework 等等，製作一個結合貼圖生成功能的聊天室。當使用者輸入需求（如：laughing）及想要的貼圖主體（如：otter）後，我們的模型會即時生成數種不同貼圖供選擇，且讓使用者可以直接點按發送至聊天室中。又因擴散模型具有每次生成結果不盡相同的特性，故相同的輸入會產生不同的結果，使每一個使用者傳送的貼圖都是獨一無二的存在，為生活增添趣味。

1. Introduction

1-1. 研究動機與目的

以日常生活經驗作為構想的出發點，我們注意到「貼圖」在現今社群媒體蓬勃發展的時代，擁有極為重要的地位。然而，就算購買或下載了許多貼圖，「找不到當下想用的貼圖」卻依然是許多人的困擾。雖然 LINE、Messenger 和 Instagram 等社群媒體大多有內建的貼圖搜尋功能，但皆為一對一的對應關係，即該貼圖被直接給予特定的標籤。因此，若擁有的貼圖不符合當下的情緒，使用者便只能放棄以貼圖傳達最直接而真實的感受。

我們認為圖像生成式模型是解決此一問題的方法。此類型的模型在使用者輸入提示詞 (Prompt) 後產生相對應的圖片，且每次的生成結果皆不同。若能將此種文字生成圖片 (text-to-image)的概念使用在貼圖生成，便能打破固定資料的限制，讓使用者真正找到「想用的貼圖」。

因此，我們希望透過即時圖像生成式模型的訓練，建立貼圖生成模型——使用者輸入需求（如：laughing）後，由模型迅速生成數種不同風格且具有生動表情的貼圖，再讓使用者選擇使用。

1-2. 文獻探討

近年來，生成式人工智慧 (Generative AI) 發展迅速且受到眾人重視，而在圖像生成的領域除了 DALL-E 和 Imagen 等模型外，由 StabilityAI 推出的 Stable Diffusion 系列開源模型亦擁有相當良好的表現。

前述模型皆利用了擴散模型(Diffusion Model)的架構；其概念源自於非平衡熱力學 (nonequilibrium thermodynamics) [1]，透過 variational inference 的方式訓練參數化的馬可夫鏈 (parameterized Markov Chain)；其中過程分為 forward process 與 reverse process，前者於影像上添加高斯雜訊 (Gaussian Noise)、後者則從雜訊一步一步回復出圖片。

以第一個能產生高品質影像的擴散模型 Denoising Diffusion Probabilistic Models (DDPM) [2]為例，Fig. 1-1中 x_T 為高斯雜訊、x_0 則為乾淨的影像；$q(x_t|x_{t-1})$為添加雜訊的 forward process、$p_\theta(x_{t-1}|x_t)$則是 reverse process。值得注意的是預測 x_0 並不是唯一的訓練方式，我們也可以透過預測 noise 或 score function 來達成相同的目標。

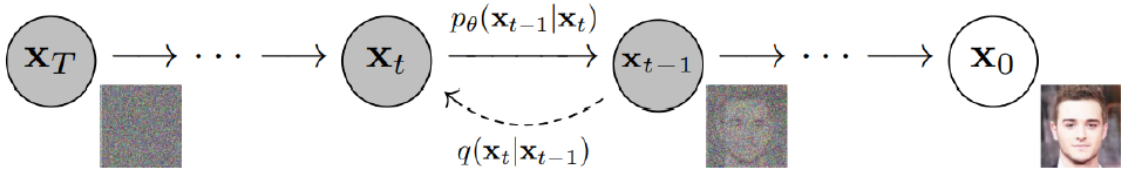

Fig. 1-1 Graphical model of DDPM [2]

然而因為 DDPM 在執行 reverse process 時需要進行多次 Markov chain simulation，使其 inference time 較長，故 Denoising Diffusion Implicit Models (DDIM) [3] 的架構被提出，其藉由加速採樣使模型可以更快速將圖片生成。

在 Stable Diffusion 系列模型中，於2023年底推出的 SDXL-Turbo [4] 提出了 Adversarial Diffusion Distillation (ADD)的構想——同時採用 Adversarial Training 和 Score Distillation 來減少 inference steps，故進一步將生成圖片所需的時間再縮短。其架構利用三個模型——ADD-student（最終訓練出來的模型）、Discriminator（辨別圖片是否由 ADD-student 所產生）、ADD-teacher（預訓練好的模型，如 SDXL）。

考量生成式模型參數量龐大（以 SDXL-Turbo 為例，其擁有約 35 億個參數），若要重新訓練模型以調整風格或用途，需要大量時間與記憶體空間。因此，Low-Rank Adaptation (LoRA) [5] 的想法被提出。此方法使我們在訓練時，不需要重新訓練全部的模型參數，而是僅訓練兩個部分——利用 rank decomposition matrix 降低模型複雜度。雖此方法最初多使用於大型語言模型的訓練，但後來在圖像生成式模型的研究上亦有很好的表現。

2. Research Methodology

我們的目標為利用圖像生成式模型即時生成可以運用於日常生活中的貼圖，故接下來分為「模型訓練」和「聊天室實作」兩大主題進行研究方法的說明。

2-1. 模型訓練

考量以貼圖的「即時」生成做為目標，在研究過各種圖像生成式模型後，我們選定 SDXL-Turbo 作為訓練對象，並以 Low-Rank Adaptation (LoRA) 的方式針對不同風格、表情和語句調整模型。模型訓練、驗證與測試的過程使用到 Pytorch、Transformer、Accelerate 和 Huggingface 的 Diffusers 等函式庫。

我們使用 LINE 的貼圖作為 Dataset 來源；資料預處理的過程包含將貼圖上的文字移除以避免提取錯誤的特徵、調整尺寸為正方形、並給予每張貼圖不同標籤（包含情境、主體、創作者名稱和我們模型的提示詞，如：happy, laugh, tiger, Bu2ma, dm-sticker）。如 Table 2-1 所示，我們採用四組 Dataset 進行了四種不同風格的模型訓練，每組 Dataset 至少使用 80 張貼圖，使模型能在訓練後生成與訓練資料風格一致的貼圖。

Table 2-1

Training Dataset

系列	我不是胖虎	肥嫩水獺	ㄇㄚˊ幾兔	胖才可愛
創作者	胡創/ 不二馬大叔	1g.wu	YUKIJI	胖才可愛
貼圖數量	80	113	118	148

(Reference: LINE STORE, https://store.line.me/home/zh-Hant)

訓練模型時，我們隨機水平翻轉 Dataset 的貼圖，以進行資料增強(data augmentation)，而我們亦使用 Adam 優化訓練成果。在訓練參數 (hyperparameter) 的選擇上，一個 batch 約 10 張左右的圖片和 1E+04 附近的 learning rate 能夠產生符合我們期望的結果。此外，為了減少一些不必要且雜亂的線條或五官變形缺失等問題，我們亦給予 negative prompt（如：bad anatomy, bad hands, missing fingers, extra eyes, low quality 等）。在 scheduler 的選擇上，我們多採用 cosine scheduler，以改善圖片資訊破壞過快的問題 [6]。

此外，為了提高模型的風格多樣性與靈活度，我們亦採用 Huggingface 支援的 Fuse Lora 混用不同風格模型；使模型在產生新風格同時，有更穩定的表現。與其他混合 LoRA model weights 的方法相比，此方法所需要的記憶體較少，且 inference speed 較快。

2-2. 聊天室實作

為了展示模型於現實生活中的應用，我們實作了一個整合聊天室以及貼圖生成與傳送的聊天室。使用者可以透過輸入當下想表達的情緒或語句（例：Sad、Congratulations!、Sorry）以及貼圖的主體（例：rabbit、otter），得到符合需求且多樣化的貼圖。聊天室的設計架構主要分為前端、後端與貼圖生成三個部分。

2.2.1 前端：

使用 HTML＋CSS 設計前端介面，並導入 Tailwind CSS 加速開發效率。JavaScript 則將前端的訊息或貼圖生成的要求傳送到後端，並且將後端生成的貼圖與其他使用者傳送過來的訊息顯示於聊天室介面上。

2.2.2 後端：

使用 Python＋Flask＋Flask-socketio 作為聊天室後端的架構：Flask 處理前端發送的要求，並傳送對應的資料、網頁回到前端；Socketio 則處理聊天室的訊息傳送。

2.2.3 貼圖生成：

使用 Huggingface Diffusers 的 Pipeline 將 SDXL-Turbo 載入後，嵌入我們訓練出來的 LoRA weights 以產生貼圖。此外，為了使貼圖傳達意思更清楚，我們亦將部分生成的貼圖加上文字——先分類使用者輸入的語句後，再將對應到的單詞顯示於貼圖。

後端收到使用者的請求（包含想表達的語句及貼圖主題）後，會判斷使用者輸入的語句為句子或單個單字；若使用者輸入句子，為了方便產生貼圖文字，我們使用 Zero-Shot Classifier 利用 Huggingface Transformers 與 bart-large-mnli[7]）協助判斷句子對應的情緒和語意。Table 2-2 列出我們目前使用的預設情緒與語意。

Table 2-2

Default expression and emotion

Happy	Sad	Angry	Tired	Hello
Congrats	Thanks	Okay	Sleep	Curious

經過 classifier 後，我們將原本的提示詞加上預設的 model 提示詞（dm-sticker），與減少五官不全的 negative prompt；再隨機擇兩個風格的 model 做 Fuse LoRA，送入 diffusers pipeline 中。最後產出的部分貼圖加上 classifier 對應的文字，透過 API 傳回前端介面；若不添加文字，則是將生成的貼圖直接回傳。

此外，雖然我們的模型在 8～12 steps 內可以生成品質良好的貼圖，但於實驗室伺服器初架設聊天室時，我們認為貼圖生成速度仍然不如預期。因此，我們分析貼圖生成的步驟與其個別所需時間，並注意到最花時間的部分並不是 inference time，而是最開始載入 SDXL-Turbo 的時間，其佔據整個貼圖生成過程超過 1/3。故我們由此著手，並發現每次 inference 都需要將存放於硬碟的模型載入使此步驟所需時間較長。因此，我們使用 deepcopy() 的方法，將存放在 GPU 記憶體的 SDXL-Turbo 直接在 memory 中複製一份做 inference；利用 GPU 記憶體傳輸速度遠高於 SSD 的特性，大幅減少此過程所需時間，使生成四張貼圖的時間由約 20 秒降至 12 秒內（下降 40% 以上），即不到 3 秒即可生成一張貼圖，真正做到貼圖的「即時」生成。

3. Experimental Results

3-1. SDXL-Turbo 以 LoRA 進行單一風格訓練

下述我們以肥嫩水獺系列貼圖訓練出來的 LoRA Model，分別就使用實例、結果多樣性與風格訓練成果，進行實驗成果的相關說明與舉例。

3.1.1 模型訓練前後比較

Fig. 3-1 使用兩組不同的情境作為範例，展示進行 LoRA training 前後的結果差異。從中可以觀察到我們模型生成的貼圖在符合 Prompt 的同時，具有風格一致性且更貼近貼圖樣式，而非單純的圖片。由此可見，我們的模型能夠產生具有特定風格且符合使用者需求的貼圖。

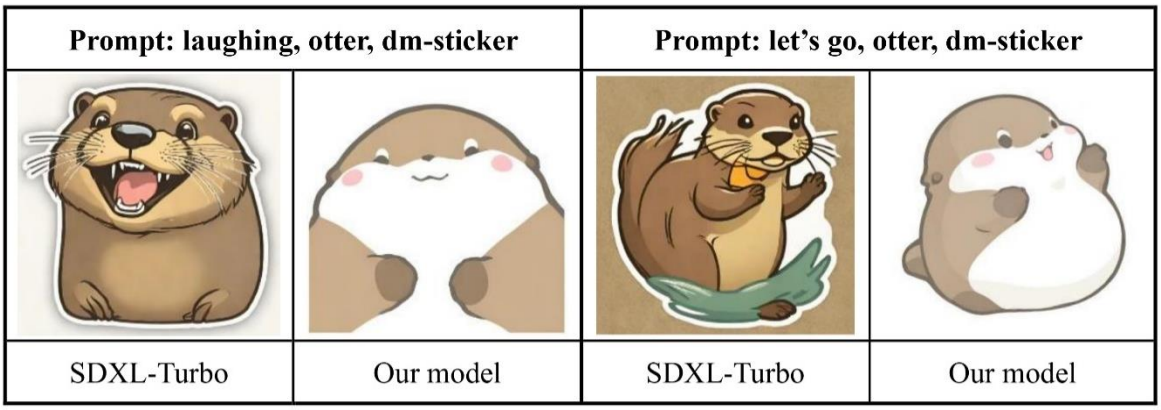

Fig. 3-1 Comparison of results generated by SDXL-Turbo and our model

3.1.2 表情、語句與情境範例及 LoRA model 多樣性展現

Fig. 3-2 呈現我們以日常生活中使用到的表情、語句或情境作為 prompt，所產生的貼圖成果，以展現我們模型於日常生活使用貼圖的表現。

此外，因為 Diffusion Model 在 inference 與 training 時皆會從 normal distribution 進行 sampling ，進而使 Diffusion Model 就算接收完全相同的參數與 prompt，每次產生的結果仍會有所差異。因此 Fig. 3-2 中我亦們可以觀察到給予相同參數與 prompt 後，每次產生的貼圖展現不同特色，但仍有維持風格一致，亦傳達出欲表達之涵義。由此可見，我們模型產生的成果具有一定程度的穩定性與多樣性，使其適用於貼圖的產生。

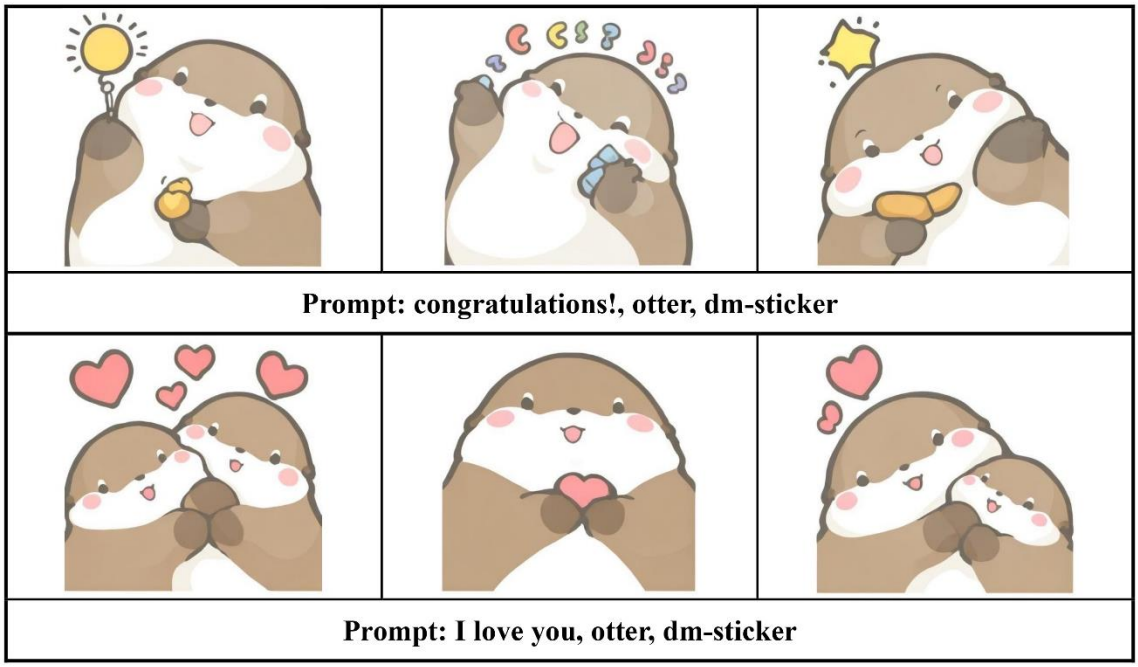

Fig. 3-2 The Stickers Generated by Our Model

3-2. SDXL-Turbo 以 Fuse LoRA 進行多種風格融合

為了提升成果的多樣性與穩定性，我們利用 Fuse LoRA 的方式，針對訓練出來的 LoRA model 進行風格混合。Fig. 3-3 呈現出混合兩種不同 LoRA model 的成果。從中可以觀察到融合後，生成的貼圖在保留原先兩種風格特徵的情況下，產生創新的風格。此外，我們觀察到 Fuse LoRA 後的模型產生結果為穩定。由此可見，在模型數量有限的情況下，我們可以利用少數幾種風格，混合出多種新風格，產生更多樣化的貼圖供使用者做選擇。

Fig. 3-3 Demonstration of How Fuse LoRA Works

3-3. 聊天室實作成果

首先，使用者進入登入畫面（如 Fig. 3-4 a 所示），在輸入名字和 Room number（若有已存在的聊天室）後，即會進入聊天室主頁面（如 Fig. 3-4 d 所示）。

若要在聊天室傳送貼圖，使用者可以藉由點按輸入框旁的「笑臉」按鈕，貼圖生成的介面即會出現。此時，使用者輸入想表達的內容，生成的貼圖會即時回傳，讓使用者藉由點選貼圖，將貼圖傳到聊天室。而若有特定想使用的貼圖主體動物，亦可以輸入以得到更貼近者需求的貼圖（如 Fig. 3-4 b 所示）。

當使用者在貼圖介面輸入句子時，內建的 classifier 會把句子對應到已經設定的情緒、語句分類中，於其中兩張貼圖上添加相對應的文字（如 Fig. 3-4 c 所示）。

最後，利用上述的聊天室介面，Fig. 3-4 展現了我們模型於日常生活中的使用，而下方連結是我們的 Demo Video：
https://drive.google.com/file/d/1CmvRdDHmZvmBVmZvrfXb2K3fUMQa6xNs/view?usp=drive_link

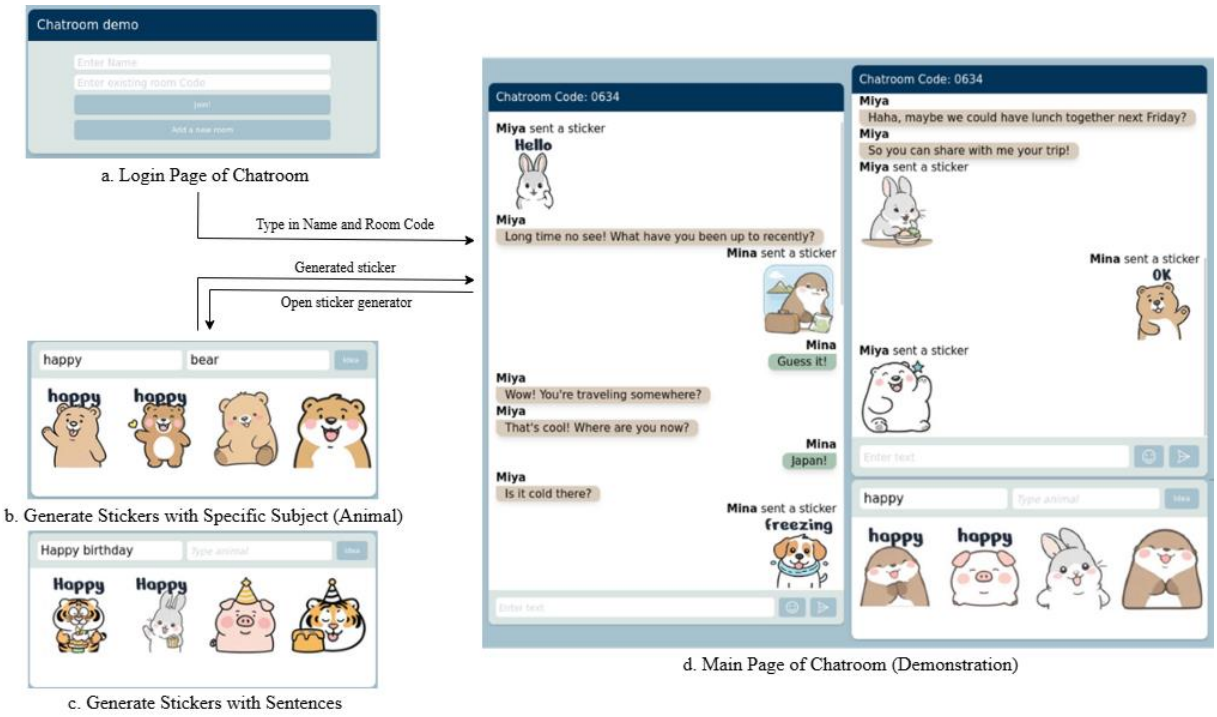

Fig. 3-4 Chatroom Demonstration

4. Conclusion

我們利用 Low Rank Adaptation (LoRA) 調整開源擴散模型 SDXL-Turbo ，以達成將圖像生成式模型用於日常生活貼圖產生的目標。從實驗結果可以觀察到我們的模型無論在產生結果的穩定性，或是多樣性皆有很好的表現；而在風格呈現的部分，其亦有成功從 Training Dataset 中學習。

此外，我們所製作的聊天室呈現出我們模型在日常生活中的實際應用——能完整表達日常對話且易於使用。由此可見，圖像生成式模型適用於此一使用情境，且此主題的應用具有未來持續發展的潛力。

5. Reference

[1] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. 2015.

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. 2020.

[3] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. 2021.

[4] Axel Sauer, Dominik Lorenz, Andreas Blattmann and Robin Rombach. Adversarial Diffusion Distillation. 2023.

[5] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang,

Lu Wang and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. 2021.

[6] Alex Nichol, Prafulla Dhariwal. Improved Denoising Diffusion Probabilistic Models. 2021

[7] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov and Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. 2019.

6. Reflection and Thoughts

這次的專題以現今流行的 Diffusion model 出發，研究其延伸的應用。由於 Diffusion model 集結了熱力學、隨機微分方程式（stochastic differential equation）等概念，因此在一開始研究論文時確實較為吃力。而進行 LoRA Training 時，為了研究最適合 SDXL-Turbo 的訓練方式，我們調整了諸多訓練參數。在聊天室的實作中，則因為對 Python 的後端不太熟悉，因此花很多時間除錯與修正。此外，因為我們一直強調「即時」，故在貼圖生成的流程上亦下了不少功夫，以提供更好的使用者體驗。

感謝黃朝宗教授給予我們這個機會參與專題，在我們閱讀論文遇到困難時，幫助我們理解論文中晦澀的概念，並且在我們實際執行遇到問題時給予諸多寶貴的建議。此外，也非常感謝另外一組同學與我們互相交流所提供的協助。