國立台北大學資訊工程學系專題報告

多功能微調式 AI 繪圖生成系統

Versatile Fine-Tuning AI Image Generation System

專題組員:李育誠、陳柏逸、呂恩齊、鄭仲哲

專題編號: PRJ-NTPUCSIE-112-12

執行期間:112 年 07 月 至 113 年 06 月

一. 摘要

本計畫將以整合 AI 繪圖各項新技術作為研究目標，為設計師提供客製化的修改與創作服務。設計師可以透過 UI 介面化操作 AI 繪圖模型，並依自身需求使用。

本計畫主要使用四種模型，分別是圖像生成之穩定擴散式模型(Stable Diffusion)、控制網模型(ControlNet Model)、拖曳式微調模型(Drag Diffusion)、預設藝術畫風 LoRA 模型。

只需要輸入一張原始圖片並打入關鍵字後，接著調整模型參數就能完成圖片生成功能。除了圖片生成功能，設計師還可以視需求使用控制網模型(ControlNet)模型讓 AI 繪圖模型照著固定的形狀、姿勢進行上色生成，使用 LoRA 模型來改變圖片的繪畫風格。又或者可以使用拖曳微調功能(Drag Diffusion)對圖片中的人物進行動作、姿勢及形狀的微調，達成理想的效果。

本研究可以為設計師提供創作靈感，大幅了減少創作的心力和時間成本，客製化的參數調整也為設計師能依照客戶需求進行個別設計。

二. 簡介

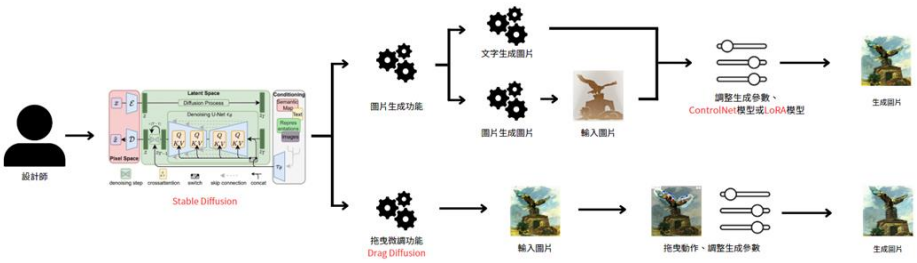
目前在繪圖委託的流程中，設計師與客戶之間的溝通存在著效率低下的問題。正常情況下，設計師需要花費許多時間來製作出一個初步的雛形，但在製作過程中無法馬上取得客戶的反饋。這可能導致設計師所提供的圖案不符合顧客的期望，進而需要進行多次溝通和修改，這樣的流程既耗時又耗力。

有鑑於 AI 繪圖技術的逐漸成熟，本計畫提出了一種新的解決方案，期待能改善這種流程。我們建議在討論階段時，利用 AI 繪圖技術快速生成客戶所想要的構圖，設計師可以透過這張圖更進一步去製作出客戶所想要的結果。如此一來可以減少設計師與客戶來回溝通上的時間成本，同時也能減少設計師修改圖案的次數，減少因為客戶不滿意而導致不必要的工作量。

三. 專題進行方式

3.1 研究方向

本計畫為了要讓設計師能更方便使用，預計使用多種新興的 AI 繪圖技術。為了確保 AI 繪圖模型所生成的圖片品質，且生成的過程中也需要確保生成圖片的構圖與輸入圖片一致，因此我們將使用最新的 AI 繪圖技術進行圖片生成。此外為了盡可能滿足客戶的需求，本計畫預計透過模型將輸入



圖一、系統流程圖

圖片進行風格轉換。或是提供拖曳式的操作界面，讓設計師可以輕鬆調整圖片的架構，盡可能完成客戶所有可能的需求。因此本計劃主要研究目標如下：

(1)圖片生成功能	(2)控制圖片構圖功能	(3)風格轉換功能
(4)拖曳微調功能	(5) GUI 介面	

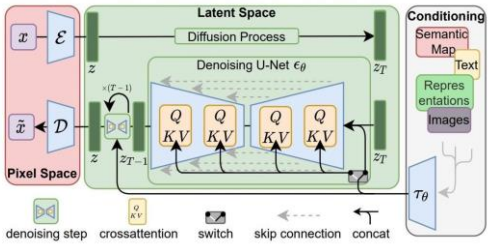
本計畫成員將分為三組，模型組、文獻組、介面組。模型組主要負責控制網模型之程式撰寫以及東西方畫家風格之 LoRA 模型訓練。文獻組主要負責查詢並實作較為創新之方法，例如拖曳微調模型。介面組主要負責整合各組之模型程式與撰寫 GUI 程式，整體系統流程圖如圖一。

3.2 文獻探討

- **擴散模型(Diffusion Model)**

擴散模型(Diffusion Model)屬於生成式模型，意味著模型可被用於生成與訓練數據類似的數據。首先透過對大量圖片連續添加高斯噪聲取得噪聲圖作為訓練數據，訓練後的模型只需輸入一張充滿雜訊的噪聲圖，模型就可以透過逐步降噪的方式取得生成影像。

根據 Ho 等學者[1]的研究，原先訓練擴散模型需要消耗大量 GPU 及時間，在單個 A100 GPU 上生成 5 萬個樣本約需要 5 天左右的時間。需要消耗大量碳足跡，高昂的硬體限制導致只有少部分人能訓練，且由於需要運行大量步驟，保存模型也花費大量儲存空間與時間。於是 Rombach 等學者[2]改善了擴散模型，提出潛在擴散模型(Latent Diffusion Model)。此方法主要是不直接在原始圖片上進行降噪，而是先將圖片透過編碼器壓縮維度較小的潛在空間(Latent Space)中再進行降噪，最後將降噪完成的潛在資料(Latent Data)輸入至解碼器後就能取得生成圖片。潛在擴散模型使訓練成本更低，且推論速度更快。本計畫預計使用的穩定擴散模型(Stable Diffusion Model)[2]也是屬於潛在擴散模型，如圖二。

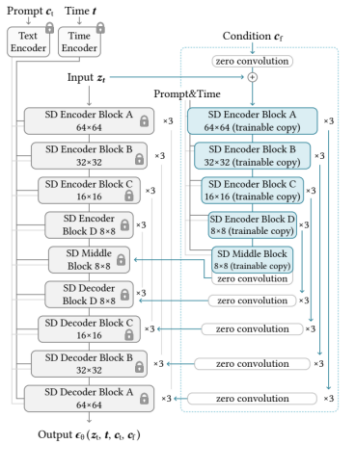


圖二、Stable Diffusion Model 結構圖[2]

- **控制網模型(ControlNet Model)**

為了確保生成圖片的構圖與輸入圖片一致，本計畫將使用由 Zhang

等學者[3]所提出的控制網模型來限制生成圖片的形狀與姿勢等構圖。當新增控制網模型到神經網路時，先將原始的神經網路鎖定，並複製一份原始神經網路的可學習複製(Trainable Copy)，接著加入可學習的零捲積(Zero Convolution)，最終神經網路與控制網模型的結果相加來獲得輸出。控制網模型在訓練前對原始神經網路無任何影響。接下來我們要使用控制網模型對穩定擴散模型(Stable Diffusion Model)[2]進行控制，模型架構圖如圖三所示。穩定擴散模型是一個基於 U-Net[4]的模型，包含編碼器區塊(Encoder Blocks)和跳連接解碼器區塊(Skip-connected Decoder Blocks)各 12 個區塊，以及中間區塊(Middle Block)共 25 個區塊。當新增控制網模型後，會透過創建 U-Net 模型的 12 個編碼器區塊和 1 個中間區塊的可學習複製(Trainable Copy)來加入控制條件。接著控制網模型中的零捲積輸出再加回 U-Net 模型的 12 個跳連接解碼器區塊和 1 個中間區塊中。



圖三、穩定擴散模型與控制網模型運作架構[2]

- **拖曳微調模型(Drag Diffusion)**

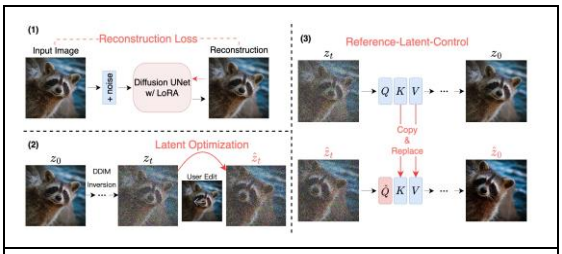
為了讓設計師對有些許缺陷的圖片進行部分調整，本計畫將使用由 Shi 等學者[6]所提出的拖曳微調模型(Drag Diffusion Model)來讓設計師可以透過對圖片拖曳拉伸點，例如讓圖片中的角色轉頭、手部姿勢抬高及物體大小微調及等拖曳重繪結果。

拖曳微調模型(Drag Diffusion Model)的運作流程圖如圖四所示，主要分為三個步驟。第一步驟為重建損失(Reconstruction Loss)，在編輯實際圖像之前，首先需要訓練 LoRA 模型對潛在擴散模型進行微調。這個階段主要是讓潛在擴散模型能更準確地取得輸入圖像的特徵，確保最後降噪後生成的圖片保持原始圖片的主體。

第二步驟為潛在優化(Latent Optimization)，是根據設計師輸入的起始點與目標點來優化輸入圖片的潛在資料。根據 Zhang 等學者[7]的研究，已經證實 UNet 模型的特徵圖中具有豐富的語義和幾何訊息，因此可以使用這些特徵來監督潛在優化過程。首先將原始圖片z_0進行 DDIM Inversion 取得潛在資料z_t，接著研究人員使用了 UNet 模型倒數第二層的特徵圖進行優化，最終將潛在資料z_t根據目標點優化成目標潛在資料$\hat{z}_t$，接著就可以將$\hat{z}_t$輸入至潛在擴散模型進行降噪。

第三步驟為參考潛在控制(Reference-Latent-Control)則是將原始潛在資料z_t和優化後的潛在資料$\hat{z}_t$分別輸入至潛在擴散模型的去噪過程中。在 Unet 模型的交叉注意層

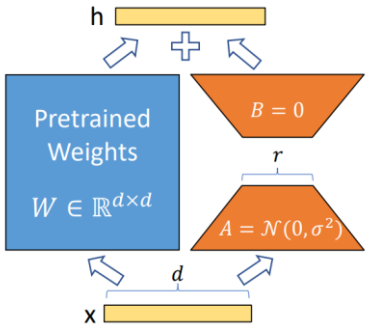
中，我們將$\hat{z}_t$所產生的 Key 和 Value，替換為z_t所生成的 Key 和 Value。透過這種方式大幅提高了原始圖像和設計師的編輯結果之間的一致性，最終完成降噪後就能取得編輯後的圖片$\hat{z}_0$。



圖四、拖曳微調模型運作流程圖[6]

- **風格轉換 LoRA 模型**

由 Hu 等學者[5]所提出的 LoRA 模型如圖五，本是用於解決大型語言模型(Large Language Model, LLM)微調所開發的一項技術。主要的做法是將預訓練好的大型語言模型權重及參數凍結，接著在 Transformer 結構中注入可訓練層(Trainable Layer)。因為不需要重新計算原始模型的梯度，因此節省相當多記憶體容量與訓練時間。

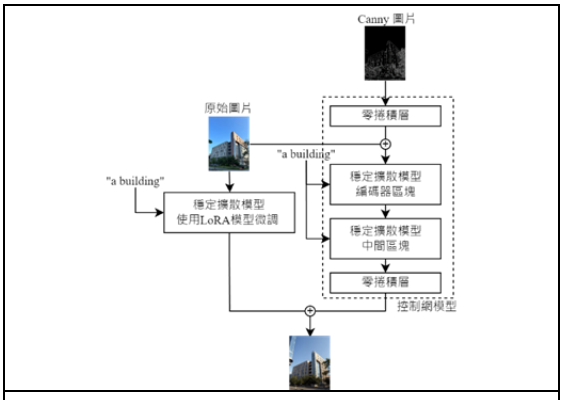


圖五、LoRA 模型微調架構圖[5]

四. 主要成果與評估

- **圖片生成功能**

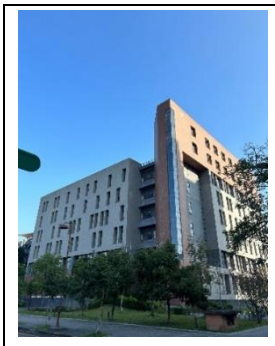
本計畫將透過整合穩定擴散模型、控制網模型及 LoRA 模型完成圖片生成任務，模型架構圖如圖六所示。進行圖片生成圖片時，設計師可以輸入一張建築物的圖片，並附上一些關鍵字(Prompt)，模型就能直接生成相應的圖片。且生成過程中會先對圖片進行預處理並輸入至控制網模型，確保生成圖片構圖會與輸入圖片一致。並透過 LoRA 模型微調穩定擴散模型完成圖片的風格轉換任務。文字生成圖片的流程大致與圖片生成圖片相同，因此不再進行詳細說明。故本計畫所設計的模型，設計師可以透過上傳圖片並簡單輸入關鍵字（Prompt），接著調整各種模型參數，就能生成獨特且引人入勝的圖像。



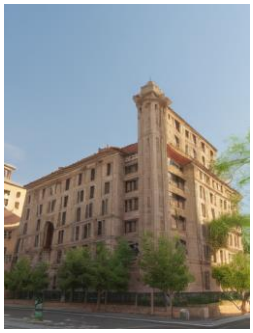
圖六、圖片生成架構圖

- **控制網模型(ControlNet Model)**

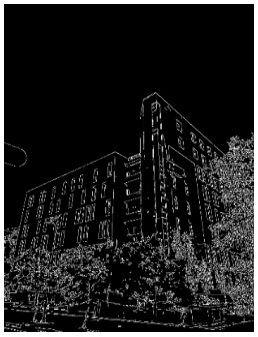
如圖七所示。我們將使用 Zhang 等學者[3]所提供的開源內容進行修改，提供了三種類型共十種不同的預訓練控制網模型，將預處理圖輸入至特定的控制網模型進行使用。預訓練控制網模型包含了：線稿類的 Canny、Lineart、Softedge、Scribble 及 MLSD，色塊類的 Depth、Normal 及 Segmentation，精細操控類的 Tile 及 Openpose。



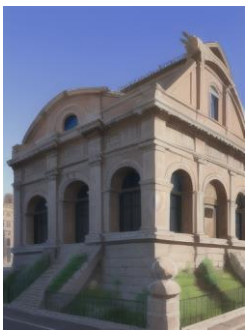
(a)原始圖片



(c)有 ControlNet 結果



(b)預處理圖

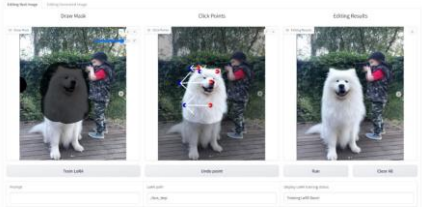


(d)無 ControlNet 結果

圖七、控制網模型對穩定擴散模型生成影響

- **拖曳微調模型(Drag Diffusion)**

拖曳擴散模型(Drag Diffusion Model)是一種新型的影像編輯方法，能夠在保持原始影像構圖的前提下，為設計師提供了前所未有的細微調整能力。模型的核心功能在於其能夠讓設計師選定影像中的特定區域進行重新繪製，同時設定多個單一方向的拖曳點。特別適合進行那些需要精細操作的調整，例如改變角色的臉部方向、調整手部姿勢或對影像中的物體進行整體的縮放，效果如圖八。



圖八、拖曳擴散模型生成範例圖

- **風格轉換 LoRA 模型**

考慮到直接對穩定擴散模型進行全微調需要消耗龐大的電腦算力和時間成本，尤其是對於訓練單一風格的畫風而言，直接全微調穩定擴散模型相當不划算。因此本計畫將採用 LoRA 模型微調穩定擴散模型，以完成圖片的風格轉換任務，不僅能夠大幅減少訓練模型所需的時間，同時也能取得良好的風格轉換效果。在設計師的建議下，我們預計會訓練東西方各三位畫家的繪畫風格，首先在西方領域，我們選擇了梵谷、莫內及雷諾瓦三位作家，結果圖如圖九；東方領域，我們選擇了張大千-水墨畫、何家英-工筆畫、葛飾北齋-浮世繪，結果圖如圖十。

五. 結語與展望

本計畫預計將各式功能整合成 GUI 介面供設計師使用，設計師僅需輸入一張圖片就能透過已訓練的 LoRA 模型微調穩定擴散模型完成風格轉換任務，且過程中使用控制網模型控制生成圖片的構圖，接著設計師也能繼續將風格轉換後的圖片輸入至拖曳擴散模型，透過拖曳調整圖片的構圖。

這樣的整合設計將使得設計師能夠更加靈活地進行圖片生成和風格轉換，從而提高工作效率並滿足不同的設計需求。

六. 銘謝

首先，我們要感謝李育誠同學，他在文獻探討、Drag Diffusion 程式實作以及專題進度規劃方面做出了重要貢獻。陳柏逸同學在 Controlnet 程式開發、LoRA 畫風訓練和競品分析項目中也展現了卓越的能力。我們也感謝呂

恩齊同學，他負責訓練資料整理、報告及 PPT 製作，為專題的順利進行提供了有力支持。鄭仲哲同學在整體 UI 介面撰寫及程式功能統整中的努力，使我們的專題達到了更高的水平。我們要特別感謝在此過程中給予我們指導和支持的老師、同學和家人。你們的鼓勵和幫助是我們完成此專題的重要力量。

(a) 梵谷　(b) 莫內　(c) 雷諾瓦

圖九、西方畫家 LoRA 模型成果圖

(a) 張大千　(b) 何家英　(c) 葛飾北齋

圖十、東方畫家 LoRA 模型成果圖

七. 參考文獻

[1] J. Ho, A. Jain, P. Abbeel. Denoising Diffusion Probabilistic Models. In Advances in Neural Information Processing Systems, Vol. 33, pp. 6840–6851 ,2020.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image ¨ synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695, 2022.

[3] L. Zhang, A. Rao, and M. Agrawala, Adding Conditional Control to Text-to-Image Diffusion Models. arXiv preprint arXiv:2302.05543, 2023.

[4] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention MICCAI International Conference, pages 234–241, 2015.

[5] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen. LoRA: Low-Rank Adaptation of Large Language Models. arXiv preprint arXiv:2106.09685, 2021.

[6] Y. Shi, C. Xue, J. H. Liew, J. Pan, H. Yan, W. Zhang, V. Y. F. Tan, S. Bai. DragDiffusion: Harnessing Diffusion Models for Interactive Point-based Image Editing. arXiv preprint arXiv :2306.14435, 2023.

[7] J. Zhang, C. Herrmann, J. Hur, L. P. Cabrera, V. Jampani, D. Sun, and M. H. Yang. A tale of two features: Stable diffusion complements dino for zero-shot.semantic correspondence. arXiv preprint arXiv:2305.15347, 2023.