國立屏東大學
National Pingtung University

不同生成模型應用於圖像翻譯任務之比較與探討

組　　員：何杰懋 (CBD109019)、莊佩蓁 (CBD109031)
謝昕諺 (CBD109023)、陳冠霖 (CBD109035)

指導老師：楊柏遠　博士

國立屏東大學
National Pingtung University

摘要

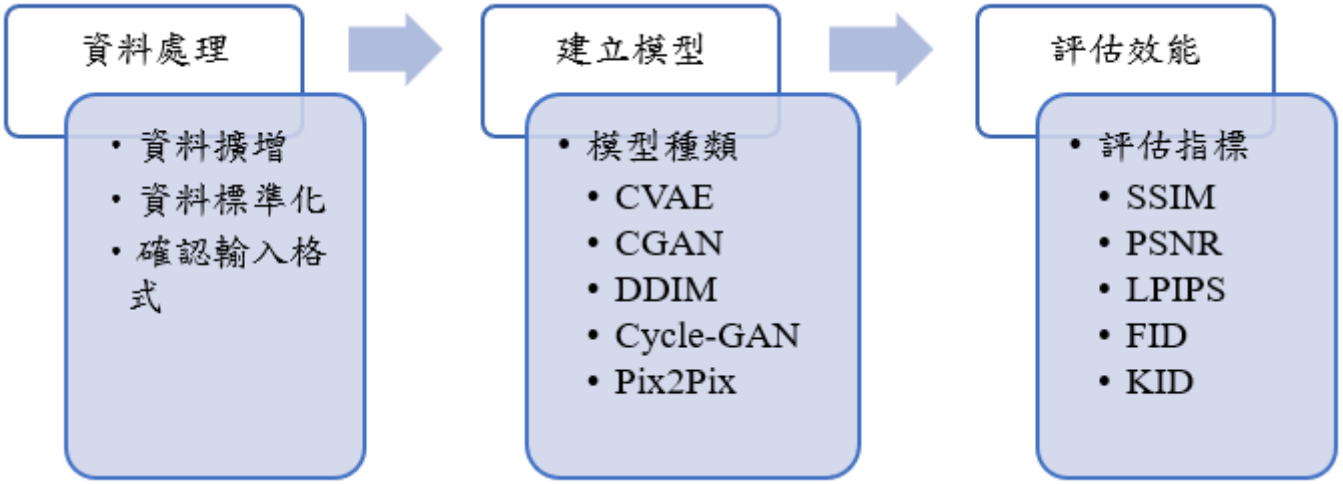
人工智慧與深度學習已然是目前科技發展的趨勢。在這之中，生成式 AI 的發展日新月異，從圖像生成、影片生成、聲音訊號生成、到聊天機器人等，都是各種不同的生成式 AI 所延伸出來的應用。生成式 AI 中又以圖像生成為最廣泛發展的領域。本專題研究分別應用不同的生成模型在圖像翻譯，並比較他們之間的效能差異；圖像翻譯是以一張圖片作為輸入，而後生成另一張不同風格的圖片。本研究使用生成對抗網路與擴散模型作為生成模型，並進一步地研究探討這些模型應用於圖像翻譯的能力以及優缺點。我們的展示頁面在https://nptuir.github.io/。

關鍵字：生成式AI、圖像翻譯、生成對抗網路、擴散模型

研究流程圖

國立屏東大學
National Pingtung University

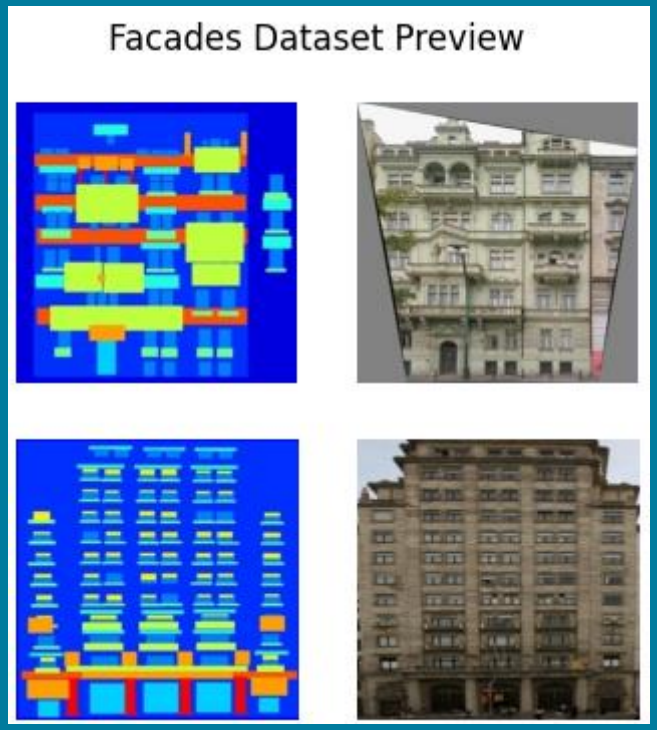
本研究旨在使用不同模型來進行影像翻譯之風格轉換任務，首先先對資料集進行處理、分析。接著分別建立五種不同的模型，這些模型分別有條件式變分自動編碼器 、條件式生成對抗網路 、去噪擴散隱式模型 、Cycle-GAN、Pix2Pix。模型訓練時也會探討訓練設定所造成的影響，並在訓練完成之後使用多種模型評估指標來判斷模型生成的優劣。接著再進一步探討各種模型及不同訓練設定下訓練成果的效能。



資料集圖片預覽

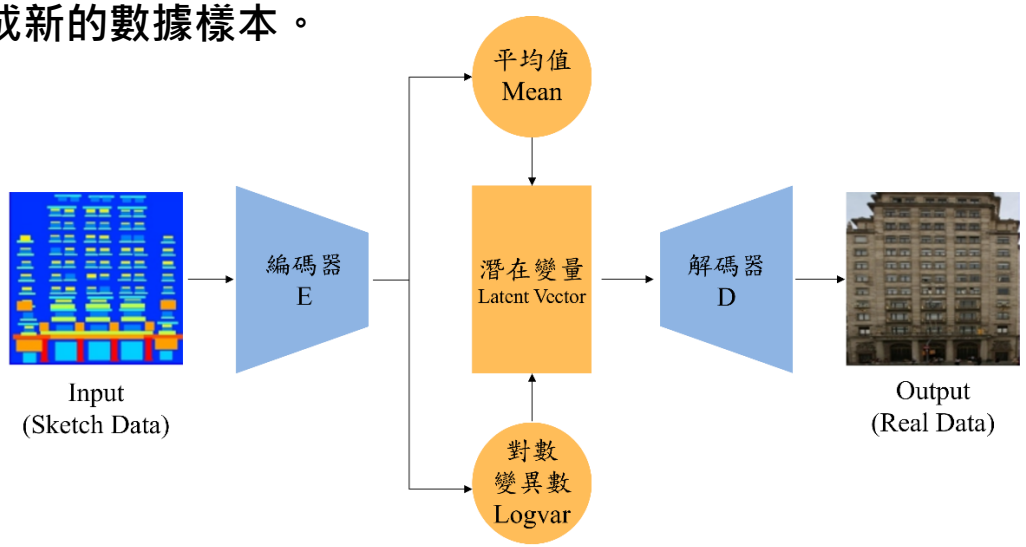
國立屏東大學
National Pingtung University

實驗中使用了資料增強，將圖片放大到300×300的解析度，再隨機從圖片中取256×256的區塊作為新圖片、以及將圖片進行鏡像水平翻轉等。將資料擴充到2400張圖片。



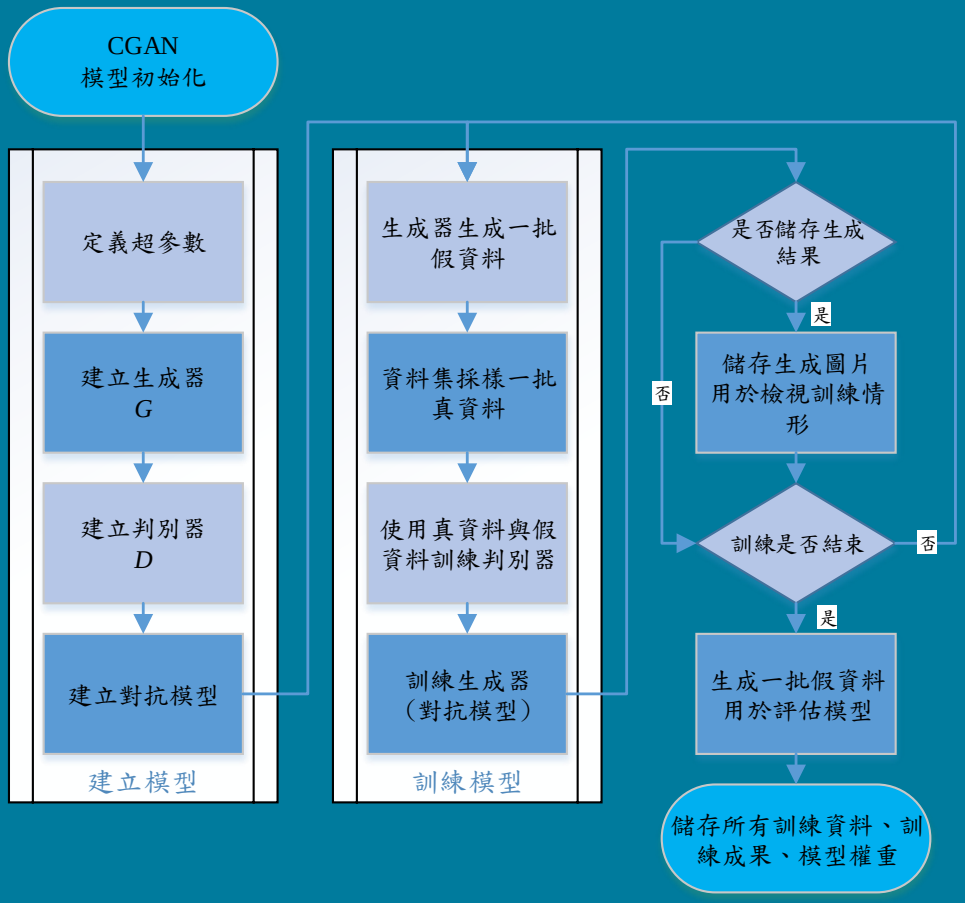
條件式變分自動編碼器 (CVAE)

條件變分自編碼器(Conditional Variational Auto-Encoder, CVAE)基本結構為編碼器(Encoder)和解碼器(Decoder)。編碼器的功能是將輸入數據轉換為潛在變數的機率分布，而在編碼器的最後一層為自定義層，負責處理潛在向量空間的部分。解碼器則是接收由編碼器生成的潛在變數樣本，然後嘗試生成與原始輸入數據相似的數據，使其可以用來生成新的數據樣本。



條件式生成對抗網路 (CGAN)

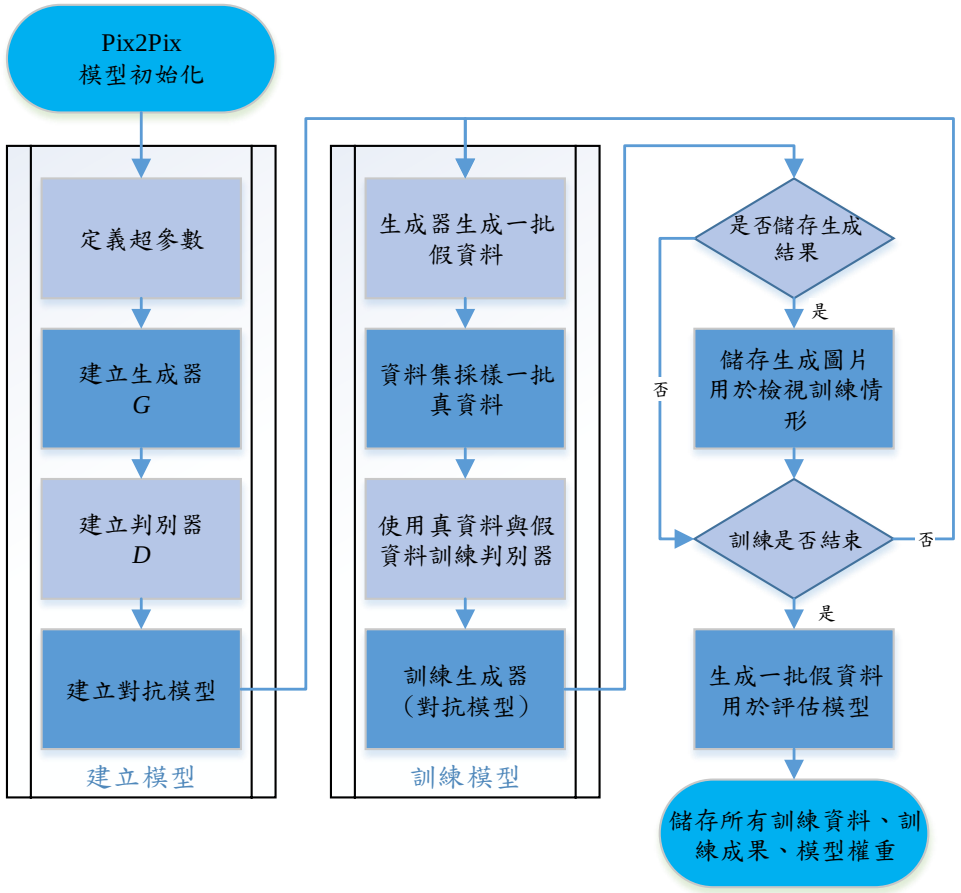
GAN的模型中有判別器(Discriminator, D)生成器(Generator, G)，CGAN與GAN的差異就是CGAN會使用條件輸入來引導生成，不像GAN生成的結果是人為不可控的。本研究CGAN的訓練流程如下圖，基本上GAN的訓練流程都大同小異。其中CGAN訓練30000次，每次訓練一批資料使用64張圖片，優化器使用Adam優化器，生成器與判別器學習率皆為0.0002，β_1為0.9與β_2為0.999。



Pix2Pix

國立屏東大學
National Pingtung University

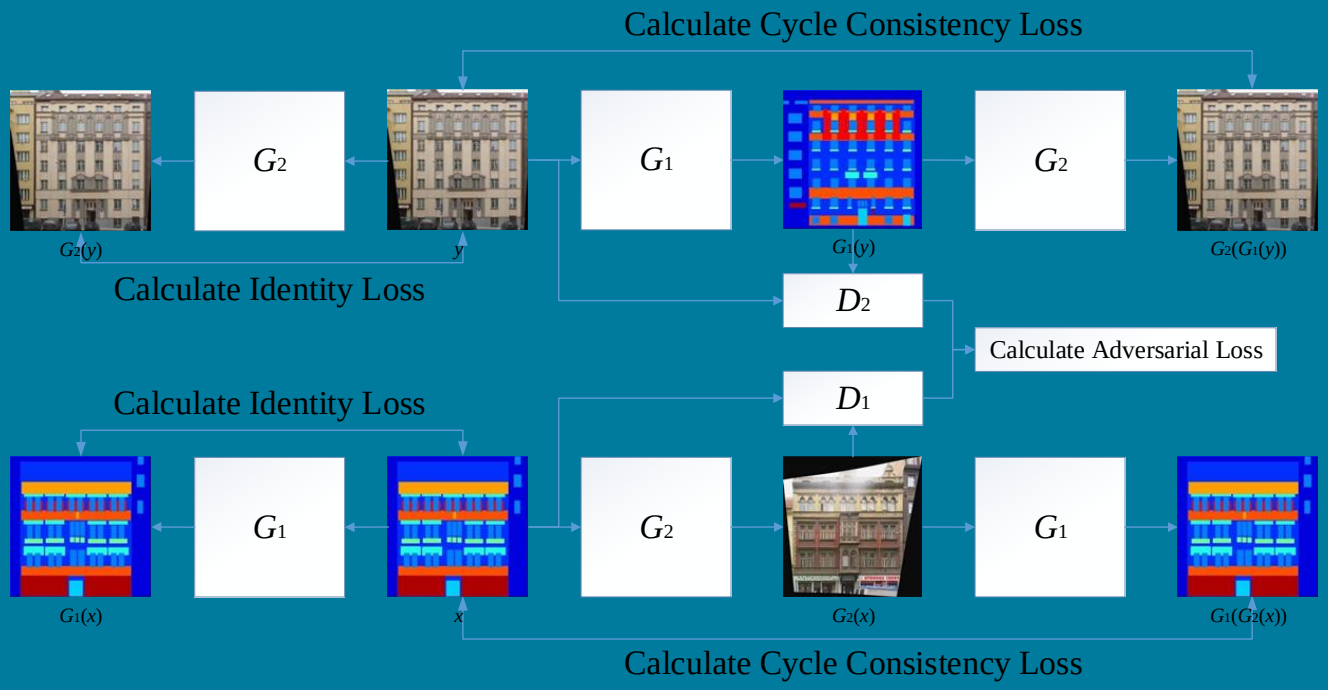
Pix2Pix是2016年被提出的生成模型，他是基於CGAN而優化改良的模型。該模型只接受一對一的資料被送入並進行圖像翻譯的任務。此模型分為兩個網路，生成器與判別器。生成器使用U-Net結構，此結構類似於AE網路，只是在下採樣層與上採樣對應層中會進行跳接，使梯度能夠傳達到前面的網路，並使下採樣的特徵與上採樣的圖片生成能夠有對應的關係。判別器使用PatchGAN架構，此架構會將輸入圖片分成許多部分，並判斷這些部分的真假，並相加以此判斷整張圖片的真假。不同於以往的GAN對於一張完整圖片判斷真假，此作法的好處是能夠更兼顧到細節的判斷。



Cycle-GAN

國立屏東大學
National Pingtung University

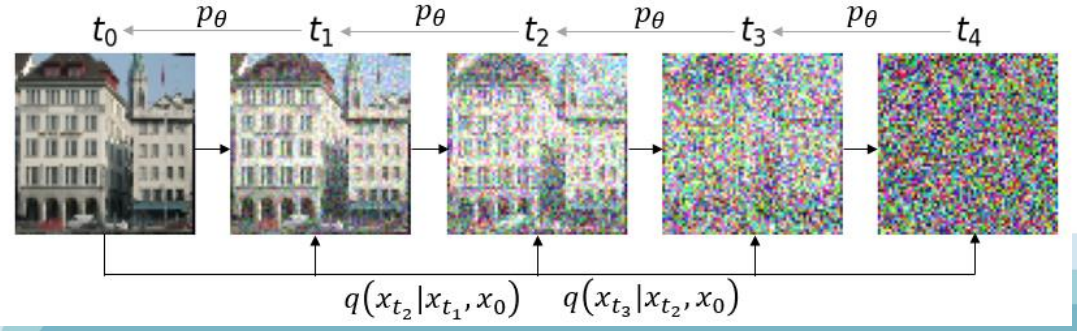
Cycle-GAN是一個GAN的另一種變化，是一種將GAN應用於影像風格轉換(Image translation)的非監督式學習演算法。與Pix2Pix不同，Cycle-GAN的輸入資料不是互相成對(pair-to-pair)的資料，即只要提供兩個不同風格的影像資料。損失分成對抗損失(Adversarial Loss)、迴圈一致損失(Cycle Consistency Loss)以及特徵損失(Identity Loss)。對抗損失是生成器嘗試生成看起來類似於某一風格的圖像，而判別器則在區分辨識生成器生成之圖像和真實圖片是否為真實圖片並計算誤差。對於來自不同風格的每個圖像都應滿足都前後向循環一致性損失，計算原影像與還原影像同位置像素間的差值。而特徵損失目的是為了保持轉換後影像的顏色組成。



去噪擴散隱式模型 (DDIM)

擴散模型是近年來非常引人關注的生成模型，其基礎為去噪擴散機率模型DDPM，擴散模型的基本原理是利用前向擴散將一張照片添加雜訊，直到圖片變成幾乎成為雜訊；接著訓練神經網路預測此雜訊並使用逆向擴散將雜訊恢復成原始圖片。前向擴散的示意如下圖。

接著為了學習到由一張雜訊的圖片返回成原圖的辦法，故需要建立一個深度學習模型來學習雜訊分布的狀況，並用於逆向擴散。本研究使用的DDIM改良了DDPM在前向擴散中使用馬可夫鏈導致採樣時間很長的缺點。做法並非使用馬可夫鏈必須要知道前一個擴散時間的分布才能得知當下擴散時間的分布，如下圖。它可以透過計算得知以初始條件t_0的圖，直接計算出特定時間步的加雜訊圖片 。



生成模型評估指標

- **SSIM**：結構相似性指數 (Structural Similarity Index, SSIM)，為比較基礎的相似度指標，其顧名思義是用於計算兩張圖片之間其結構的相似性，此指標以圖片的亮度、對比度以及結構為計算的核心，透過考慮這三個因素來計算圖片失真的程度，這個指標因為考慮結構等因素，所以計算出來的結果會更符合人類的感知。
- **PSNR**：峰值訊噪比 (Peak Signal-to-Noise Ratio, PSNR)是用來評估兩張圖片相似程度的指標。這個指標將圖片以訊號處理的方式計算其相似度，將圖片訊號以PSNR計算後的單位可視為分貝數，並依據分貝大小來判斷圖片的相似度。通常結果為30dB~50dB時圖片差異肉眼較難看出；低於30dB時肉眼可以明顯看得出圖片的不同。
- **Kernel Inception Distance (KID)**：KID是基於Inception網路的計算方式，Inception網路是一個已經訓練好的深度學習模型，其使用ImageNet資料集[18]來做學習，資料集總計有1000種物件的分類，資料量共超過一百萬張圖片。但在計算KID時只會使用其中部分網路層以得到圖片的特徵圖。KID可以透過計算Inception網路出來的圖片特徵，將生成圖片與真實圖片特徵的平均值差異之平方計算出來並衡量兩個特徵之間的差異。此外KID還有一個三次核的無偏估計值，這個估計值能夠讓計算出來的結果更貼近人類的感知。

生成模型評估指標

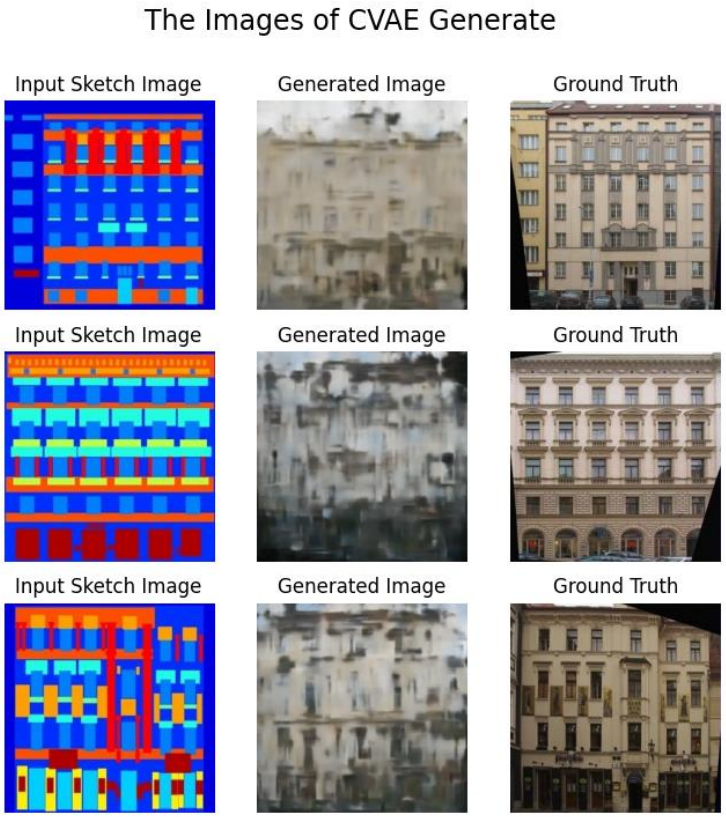
國立屏東大學 National Pingtung University

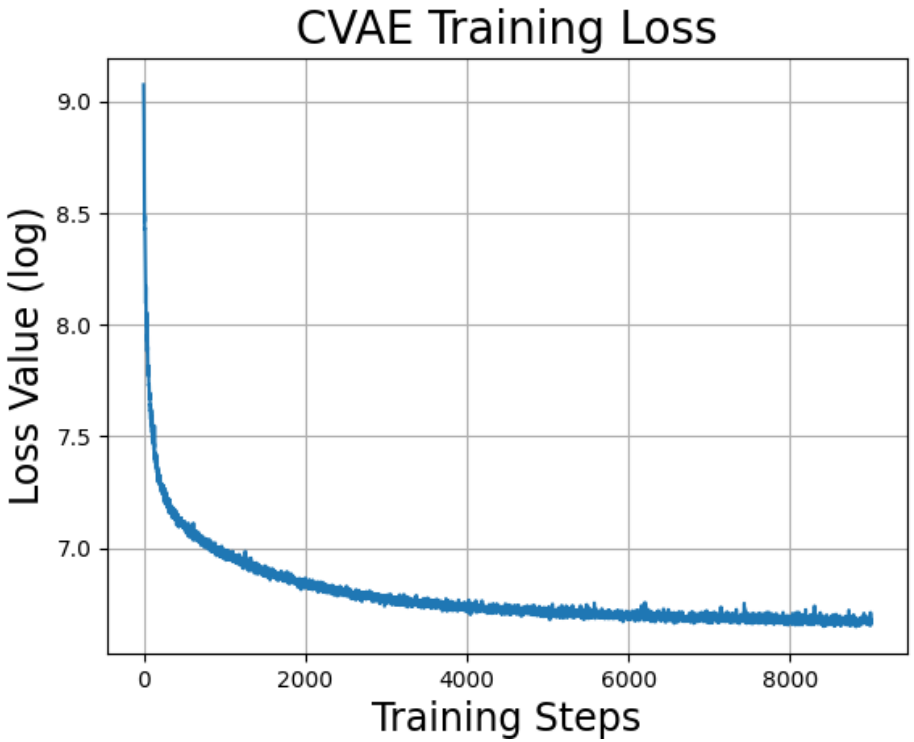
- **Fréchet Inception Distance (FID)**：FID分數是非常廣泛的圖像生成模型判斷的指標之一，其基礎也使用了Inception網路來萃取圖片的特徵。FID適合評分生成模型的多樣性，但因為模型基於特徵提取所以並不會考慮特徵位置的合理性。比起KID，FID計算量較少，所以還是比較廣泛用於生成模型的訓練，FID分數越低代表模型生成圖片之質量較佳。
- **Learned Perceptual Image Patch Similarity (LPIPS)**：Learned perceptual image patch similarity (LPIPS)為計算兩張圖片的感知損失，這個指標也是基於深度學習模型提取特徵後的特徵相似度。比起不使用深度學習的PSNR與SSIM，LPIPS也能更貼近人類的感知。LPIPS計算出來的值越低代表圖片的相似程度越高。本實驗的神經網路使用11層的AlexNet，AlexNet於 2012 年被發表。它當時在 ImageNet 資料集上展現了最先進的性能，用於LPIPS時速度最快且效能最好。

研究成果 (CVAE)

國立屏東大學
National Pingtung University

下圖左圖為CVAE對於測試資料集生成的圖片，右圖為訓練損失。

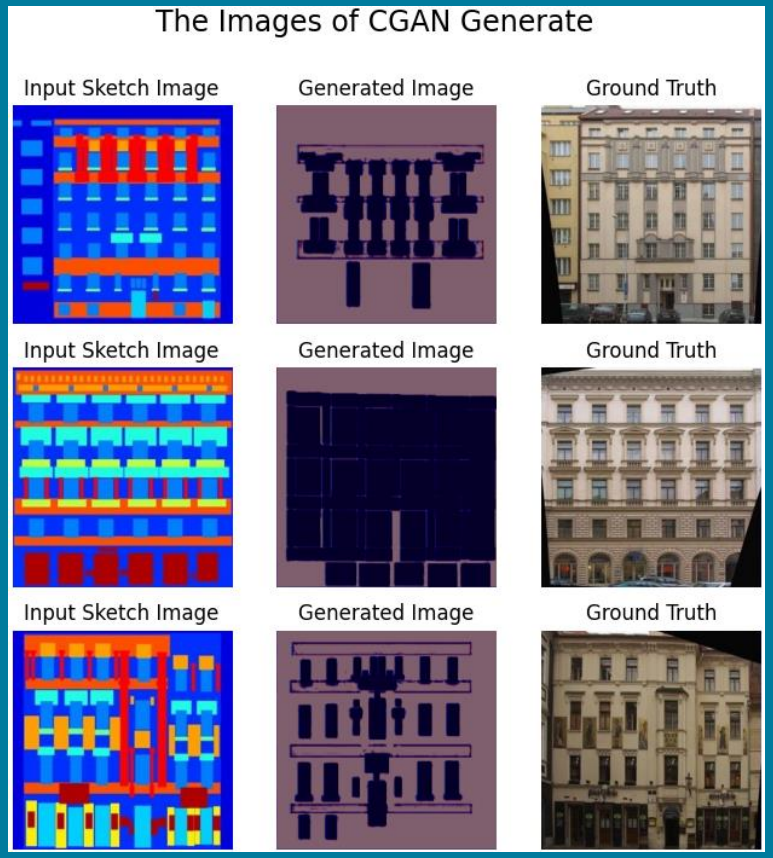


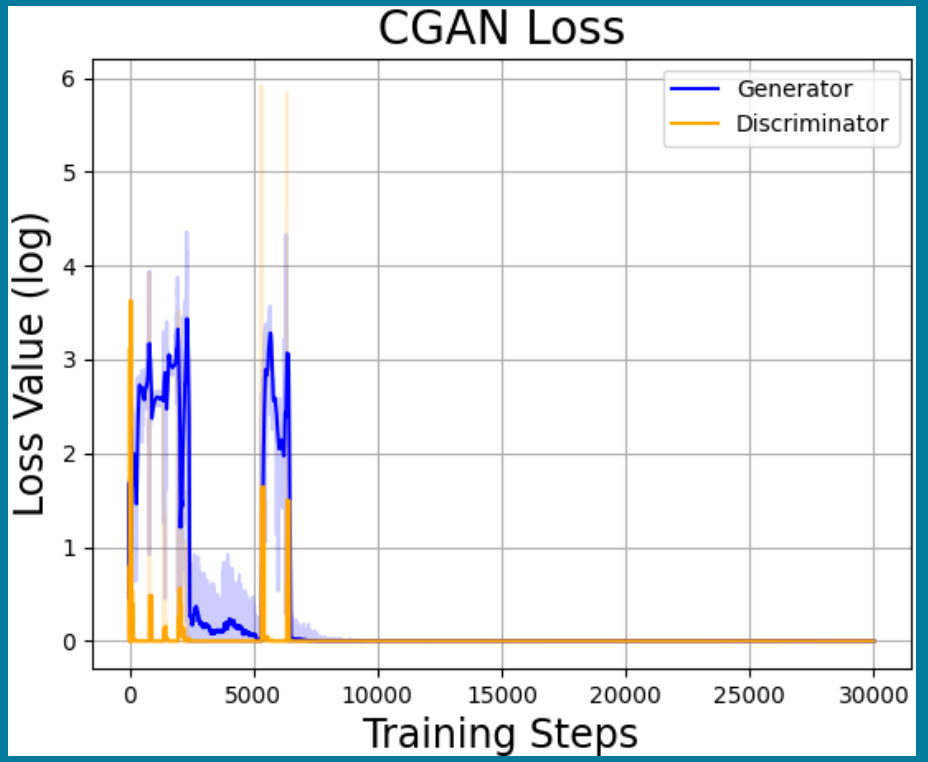


研究成果 (CGAN)

國立屏東大學
National Pingtung University

下圖左圖為CGAN對於測試資料集生成的圖片，右圖為訓練損失。

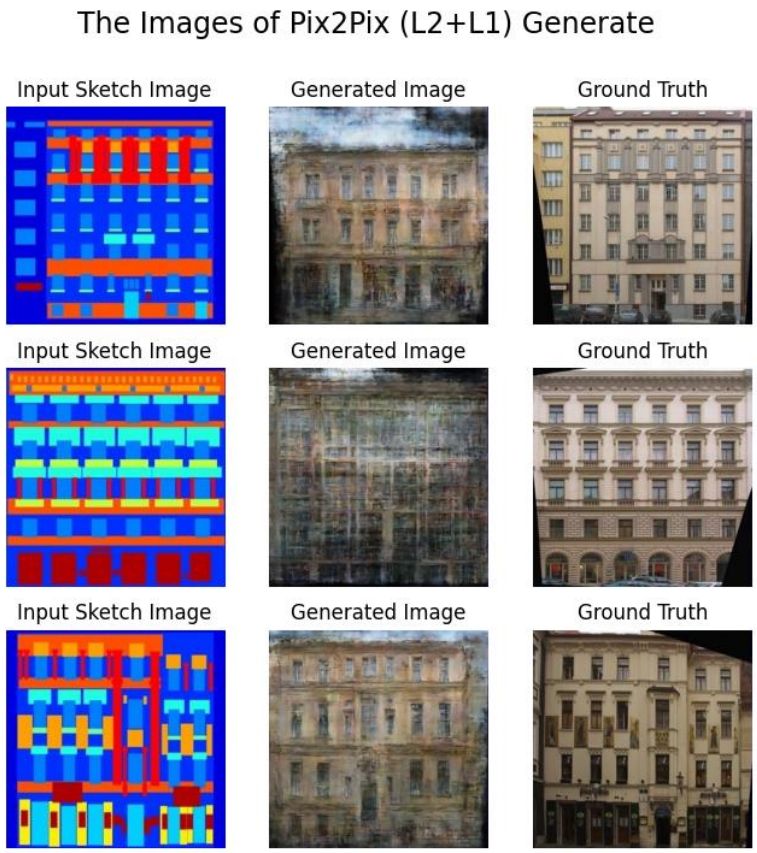


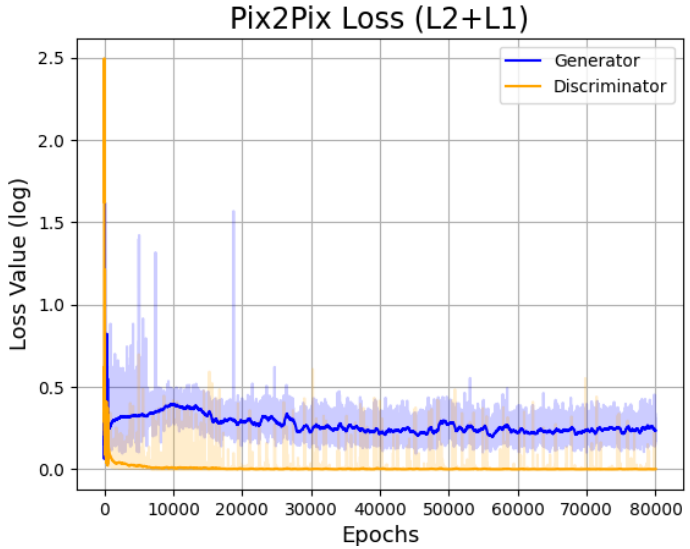


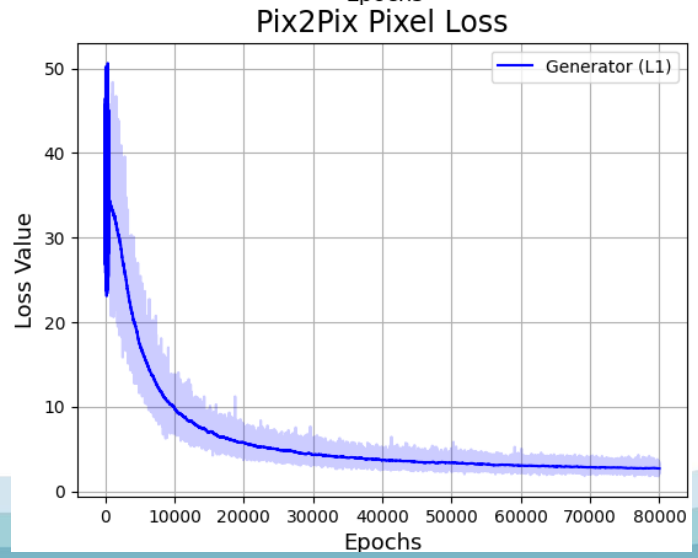
研究成果 (Pix2Pix)

國立屏東大學
National Pingtung University

下圖左圖為Pix2Pix對於測試資料集生成的圖片，右圖為訓練損失。

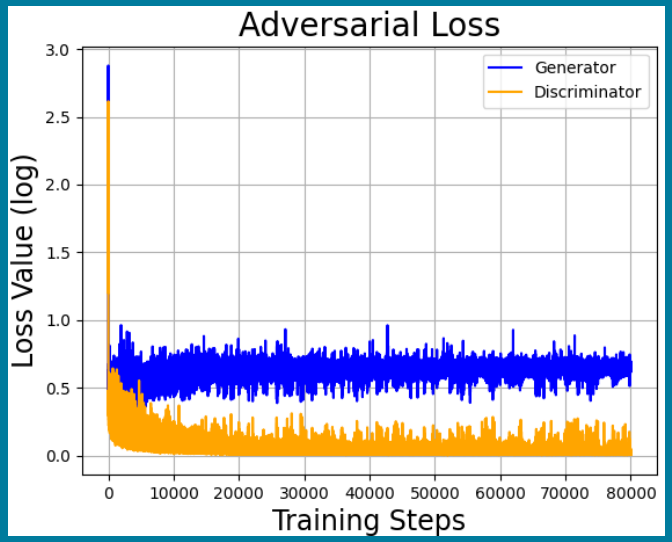


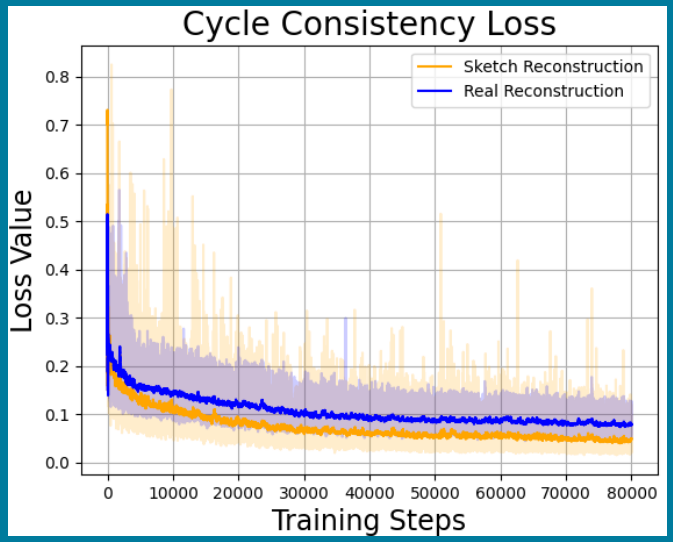


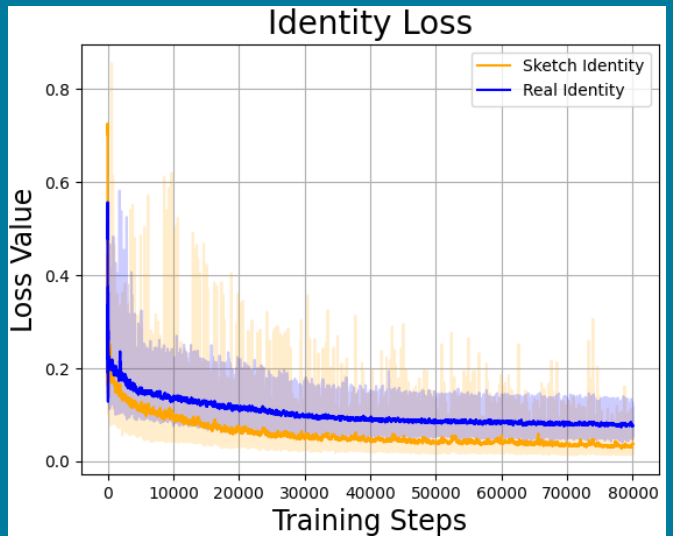


研究成果 (Cycle-GAN)

下圖為Cycle-GAN訓練時的三種訓練損失。

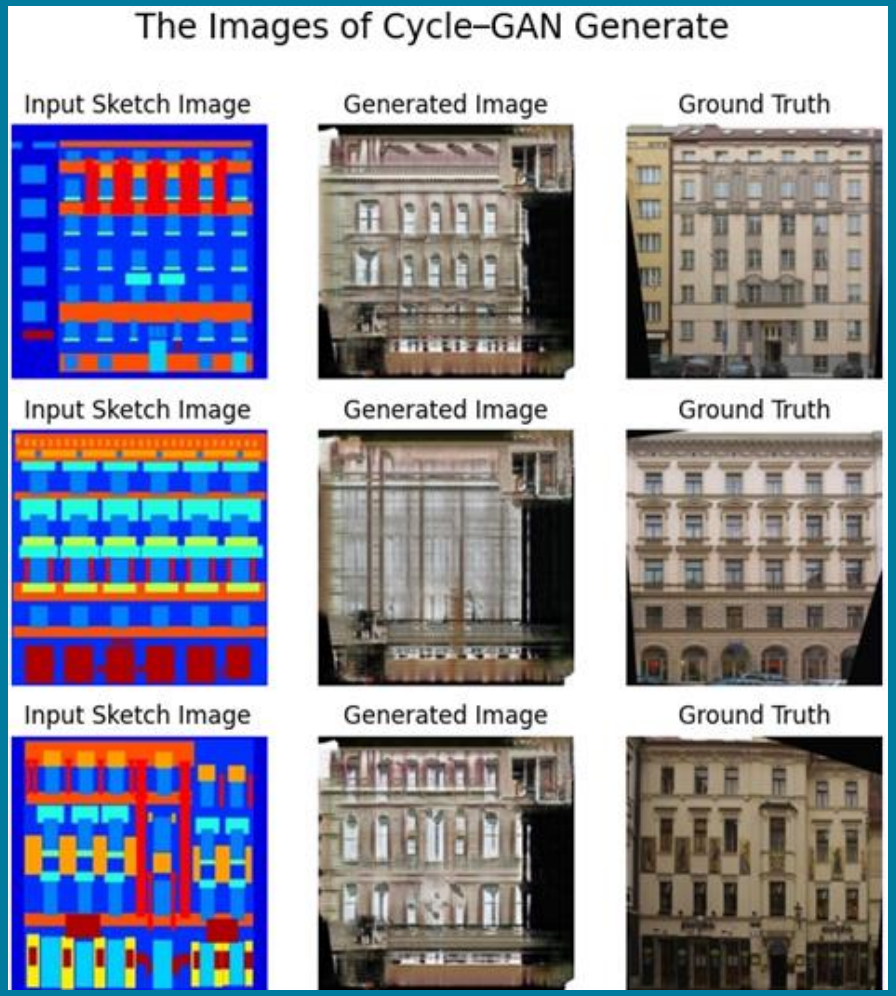






研究成果 (Cycle-GAN)

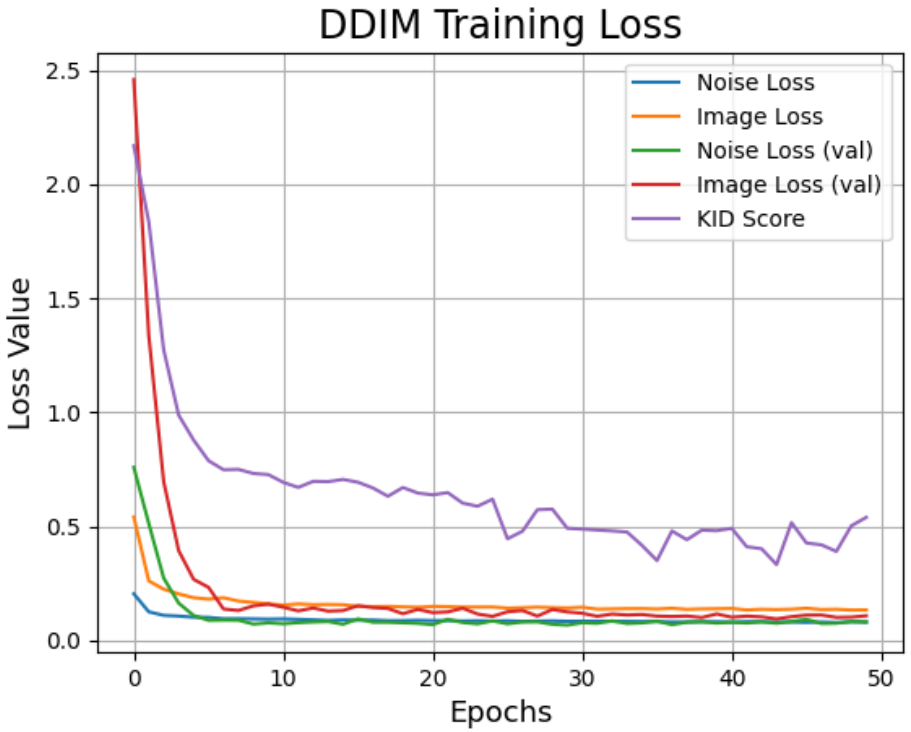
下圖為Cycle-GAN對於測試資料集生成的圖片。



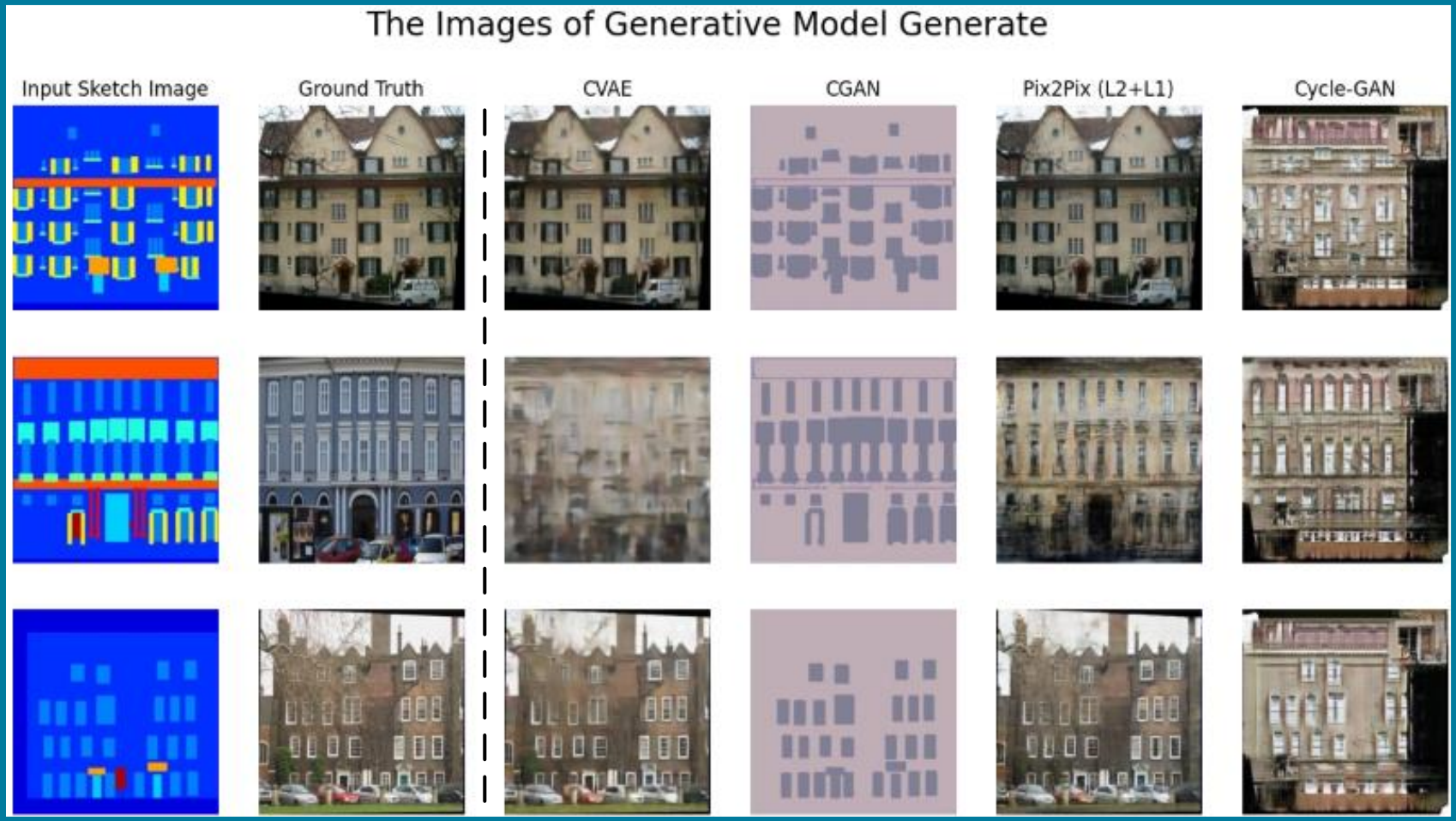
研究成果 (DDIM)

國立屏東大學
National Pingtung University

下圖左圖為DDIM生成的圖片(圖片使用64×64)，右圖為訓練損失。



研究成果比較


The Images of Generative Model Generate
Input Sketch Image
Ground Truth
CVAE
CGAN
Pix2Pix (L2+L1)
Cycle-GAN

研究成果比較

	CVAE	CGAN	Pix2Pix (L2+L1)	Pix2Pix (Bce+L1)	Cycle-GAN	DDIM (64px)
PSNR	14.387	7.492	**23.725**	13.675	9.812	2.346
SSIM	0.3	0.124	**0.588**	0.228	0.055	-0.004
FID	6.64	16.111	**4.453**	14.063	8.999	18.462
KID	0.227	0.443	**0.027**	0.329	0.14	0.264
LPIPS	0.355	0.696	**0.186**	0.431	0.375	0.439

結語與未來應用

國立屏東大學 National Pingtung University

- 本專題使用多種生成模型來執行圖像轉圖像的圖像翻譯、風格變換類型任務，研究成果顯示在訓練中Pix2Pix與CVAE在經過訓練可以得到與訓練資料及幾乎一樣的圖片，但使用測試資料生成則圖片會出現明顯瑕疵，訓練上可能出現了過度擬合。CGAN則完全無法生成可辨識的結果，模型生成結果幾乎都不可辨識，且訓練上還出現了梯度消失。而Cycle-GAN訓練後可以大致學習到訓練資料的一些表現特徵，並進行圖像風格的變換應用，且無論面對訓練資料抑或是測試資料都可以生成與草圖條件相應的結果。惟資料集數量較少故無法完整且廣泛地學習到圖片的細節特徵。
- 在需要大量運算資源的DDIM的訓練中，我們發現DDIM訓練時間比起其他模型來說更久。而且DDIM訓練上也很容易因為資料量不夠而無法良好的學習到圖片的特徵，即使使用資料增強對於訓練時資料量可能還略顯不足，未來訓練上則需要注意資料量是否充足。以及未來也需要著重研究DDIM在條件生成上的方式，何種方式能夠使擴散模型學習到更好的條件與對應風格之圖片生成，這都是未來研究值得探討的部分。
- 本專題的風格變換應用在未來也可以應用於許多不同的場合，例如在影像處理領域中可用於黑白圖片轉換為真實圖片、2D圖片轉換為3D圖片、將圖片轉換為短時間的影片等。在訊號處理領域也能根據不同風格的聲音、針對音樂等進行風格的變換。在機器人控制領域也可以用於機器人馬達控制的應用例如將機器人前進的腿部馬達控制參數改為向左走或者向右走、機器視覺轉換等應用。也能用於資料擴增等，例如分類任務中若某種類型資料不足也可以使用風格變換來補全資料不足的部分，對於一些項醫學影像等較難收集到的資料也有一定的幫助。總得來說資料風格變換的應用非常廣泛，在許多領域都能夠使用此技術來生成資料。

國立屏東大學
National Pingtung University

其他連結

專題研究成果預覽網頁

專題報告書完整文章

謝謝聆聽
Thanks for your attention!