二次元畫像條件生成對抗網路

專題編號:**111-CSIE-S027**
執行期限：**110** 年第 **1** 學期至 **111** 年第 **1** 學期
指導教授：謝東儒
專題參與人員：**108820024** 梁志榆
108820011 莊士頡

ABSTRACT

近年來虛擬主播 (VTuber) 產業正蓬勃發展，但是想要成為 VTuber 必須要擁有的模型製作價格卻非常高昂，不僅需要繪師設計角色還需要建模師把模型按照設計出的角色建立模型。本團隊希望利用條件生成對抗網路 (Conditional GAN)，使輸入僅僅為一張 2D 角色圖片以及目標動作參數，輸出為同角色且擁有目標動作與表情，大幅降低成為 VTuber 所需成本。

KEYWORDS

條件對抗生成網路, 表情生成, Face Animation

1 INTRODUCTION 導論

最近幾年 VTuber 行業迅速崛起，VTuber (Virtual YouTuber) 虛擬實況主，為真人控制著一個虛擬形象的聲音以及動作，在各種網路平台 (例如: Youtube、Bilibili、Twitch、Facebook) 上傳影片或者直播。

如果想成為 VTuber 需要做些什麼? 首先，需要一能模型讓你可以自由操縱它，常見的角色模型分為兩種:3D 模型和 2D 模型，兩種模型皆價值不菲。

本團隊受到 Khungurn[1] 的啟發，如果可以使角色設計師畫好的圖片直接動起來，不需要經過建模師之手，成為 VTuber 的門檻是否會降低許多? 若是這樣，想要擁有自己的角色並成為 VTuber 僅僅需要委託繪師幫你設計一張角色圖片，或甚至利用生成對抗網路來產生一張免費的圖片!

2 研究方法 MATERIAL AND METHODS

此系統的輸入需要一張角色圖片以及動作參數向量，圖片大小為 256x256 像素，為 RGBA 形式，且要有透明背景。說得更精確一點，若是不屬於角色的像素，則 RGBA 值為 (0,0,0,0)，若是屬於角色的像素，Alpha 值不可為 0。角色的臉必須面對照片平面，頭頂到下巴需要在正中間的 128x128 像素，並且眼睛及嘴巴必須是張開的 (此系統也可以運作在眼睛及嘴巴閉起來的角色圖片上，但是即使調整參數眼睛及嘴巴也不會張開，因為系統沒有足夠的資訊來推斷他們張開時是什麼樣子)。

如同前面提到的，總共有 6 個"姿態向量"來控制角色的動作，其中前三個用來控制臉部表情 (左右眼及嘴巴開合) 的參數介於 [0,1]。剩下三個參數控制頭部的轉動，用 3D 動畫的術語來說，角色的頭部是被兩個“關節”所控制。"頸根關節"是脖子連接身體之處，而“頸尖關節”為脖子連接頭部之處。三個參數皆沿著頸尖關節旋轉範圍介於 [-1,1]。

3 建立資料集和訓練過程

本章節我們將說明如何產生 GAN 訓練資料集，訓練資料集中如前面所提到的，共有 5 個控制向，算上左右眼的閉合共有 6 個參數，分別是左右眼閉闔、嘴巴閉合和頭部 XYZ 旋轉各 15 度，使其能做出基礎的表情。

3.1 關於訓練資料集

Table 1: 資料集數量

	訓練資料集	驗證資料集	測試資料集
模型	7,000	129	200
初始姿勢圖片	7,000	129	200
表情改變圖片	490,000	9,030	14,000
轉頭圖片	490,000	9,030	14,000
總計圖片數	987,000	18,189	28,200

我們使用爬蟲於 niconi 和 bowlroll 和模之屋使用爬蟲下載所有可合法供使用的模型並手動選出約 7500 隻較適合用於訓練的模型作為產生資料集的模型，並隨機打散區分自的用途，我們盡量使驗證和測試資料集從以上三網站下載的模型維持相同比例。

3.2 使用工具介紹:my_saba

我們使用的 my_saba 版本是由 ming173899 修改 benikabocha 所創造的 saba 而來，原始版的 saba 僅有檢視和編輯模型的功能，my_saba 版本則是加入了鏡頭調整和產生資料集的功能，而我們也調整了 my_saba 的部分功能使其符合我們的需求。

3.3 操作步驟

因為要使用自動化程式生成訓練資料集，為了要讓自動化程式可以有效的識別出需要做隨機生成訓練資料集的參數，我們需要手動把每一個模型的上述五個參數統一重新命名。

第二步為產生 mmd 模型的檔案列表，此部分我們使用 python 掃描模型檔所在位置，然後找出所有副檔名為.pmx 的檔案以換行來區隔每個檔案的位置，並存成 txt 檔。

最後一步為使用 my_saba 產生訓練資料集: 由於產生的資料集的人頭部分需要置於中間的 128x128 的位置，因此我們會使用 my_saba 紀錄各模型的 camera 位置，透過點擊每一張模型的頭頂和下巴後，my_saba 會自動將其置中並調整大小，之後儲存該鏡頭參數，並生成資料集，此部分花了我們近 1 個月。

3.4 訓練過程

本系統訓練可拆解為三個部分: 表情產生器 (Face Morpher), 轉頭器 (Face Rotator) 和合成器 (Combiner)。

- 表情產生器 (Face Morpher): 使用 Pumarola et al. [3] 的 Attention-based generator 模型，架構如圖1。
- 轉頭器 (Face Rotator): 使用 Pumarola et al. [3] 的 Attention-based generator 和 Zhou et al.[4] 的 Appearance flow 模型，架構如圖2。
- 合成器 (Combiner) 使用 Khungurn[1] 的模型，架構如圖3。

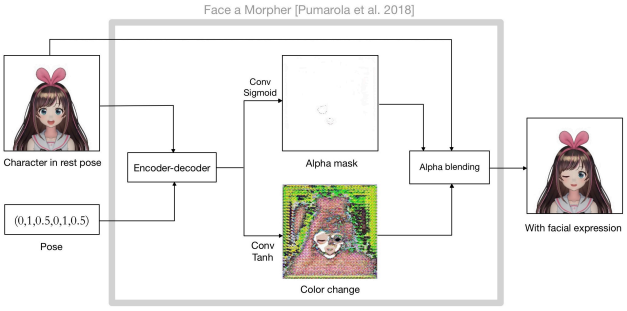

Figure 1: Face Morpher 系統架構

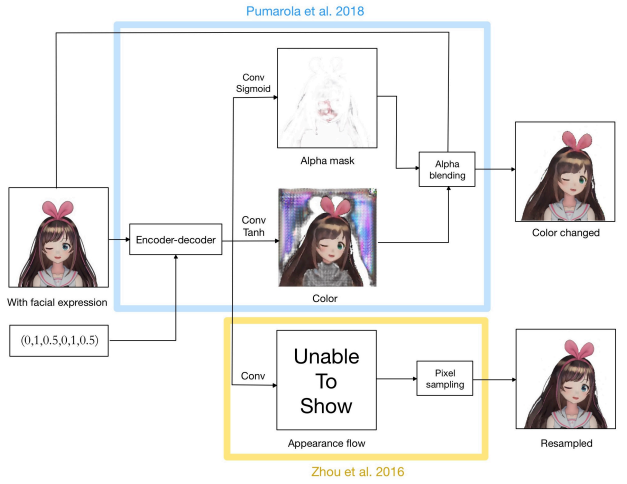

Figure 2: Face Rotator 系統架構

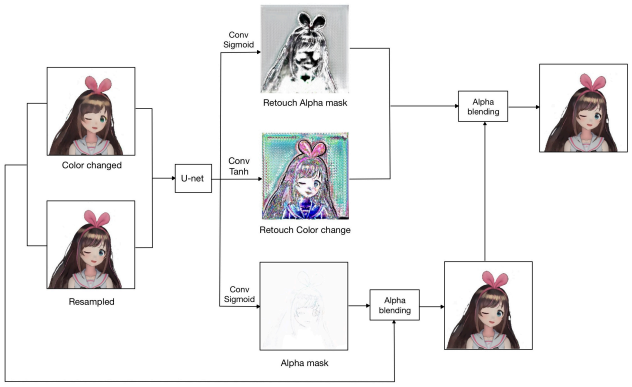

Figure 3: Combiner 系統架構

- 轉頭器與合成器之鑑別器 (Discriminator): 使用 pix2pix[2] 的鑑別器模型，鑑別器網路如圖4

神經網路層	輸入dim	輸出dim
Conv2d	7	64
LeakyRelu(negtive_slope=0.1)	64	64
Droupout(p=0.2)	64	64
Conv2d	64	128
InstanceNorm2d	128	128
LeakyRelu(negtive_slope=0.1)	128	128
Droupout(p=0.2)	128	128
Conv2d	128	256
InstanceNorm2d	256	256
LeakyRelu(negtive_slope=0.1)	256	256
Droupout(p=0.2)	256	256
Conv2d	256	512
InstanceNorm2d	512	512
LeakyRelu(negtive_slope=0.1)	512	512
Droupout(p=0.2)	512	512
ZeroPad2d((1,0,1,0))	512	512
Conv2d	512	1
Tanh	1	1

Figure 4: 鑑別器網路

4 RESULTS & DISCUSSION 研究結果與討論

4.1 研究結果

目前的結果十分讓我們滿意，效果如圖5，且性能方面於 RTX3060 上面可達成一秒生成 30 張以上，算是流暢的，但目前仍有些許高頻流失和鋸齒感的問題待解決。

(a) 系統輸入　　(b) 系統輸出

Figure 5: 輸入與輸出展示 **1** (模型來源:[5])

REFERENCES

[1] Pramook Khungurn. 2020. Retrieved may 31, 2022 from https://pkhungurn.github.io/talking-head-anime/
[2] Tinghui Zhou Alexei A. Efros Phillip Isola, Jun-Yan Zhu. 2017. Image-to-Image Translation with Conditional Adversarial Networks. https://doi.org/10.48550/ARXIV.1611.07004
[3] Albert Pumarola, Antonio Agudo, Aleix M. Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. 2018. GANimation: Anatomically-aware Facial Animation from a Single Image. https://doi.org/10.48550/ARXIV.1807.09251
[4] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A. Efros. 2016. View Synthesis by Appearance Flow. https://doi.org/10.48550/ARXIV.1605.03557
[5] 虛研社 Official. 2020. Retrieved may 20, 2022 from https://www.aplaybox.com/details/model/KPDltcb2pNNt