1. Introduction

The objective of our project is to create a talking face generation system that can produce emotional talking face videos. As emotion is a key feature in communication, we are adding facial expressions to talking faces for a more realistic representation. This system takes into account an emotional condition, a facial image, and a targeted dialogue (either in text or audio format) as inputs to generate the desired output. To achieve this goal, we developed two pipelines that generate emotional talking face videos using a single facial image, emotional condition, and targeted dialogue provided through text converted to audio by Text-to-Speech(TTS).

2. Pipeline

2.1. Our initial pipeline

The primary pipeline we suggested achieves emotion representation and lip synchronisation at once.

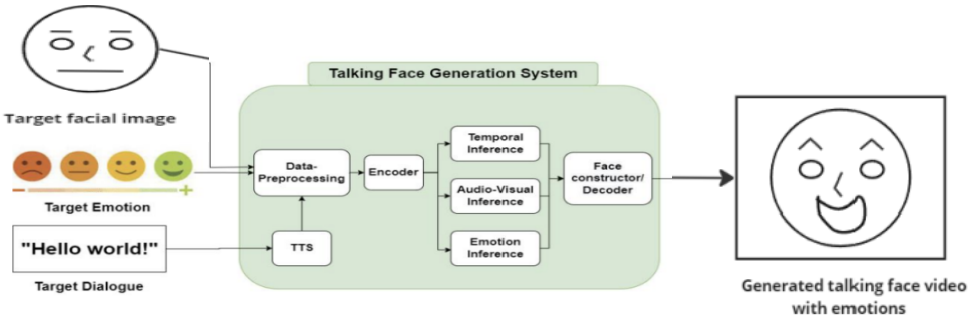

Figure 1. First pipeline

Inspiration for this pipeline was taken from the emotional discriminator presented in [1]. This introduces a model capable of generating an emotional facial video with only an image input and audio. The achievement was made possible by incorporating an emotional discriminator, which essentially functions as a video emotion classifier, offering valuable feedback to the generator. Additionally, LSTM layers were integrated into the generator, serving as an emotional encoder to comprehend the distinctive features of all six of Ekman's categorical emotions (anger, fear, disgust, happiness, sadness, and neutral). For training their model, they utilised the CREMA-D [2] emotional video dataset, which is an open-source collection containing recordings of 91 actors delivering 12 sentences for each of Ekman's six emotions.

2.2. Our final pipeline

Due to the visual and emotional quality being very unrealistic, in contrast to the first pipeline where both lip synchronisation and emotional generation are integrated into a single model, the second pipeline adopts a different approach by separating these tasks into two distinct steps, each with its own dedicated model. This process is visually depicted in Figure 2 and involves the following stages: "pre-processing (emotion generator)" and "talking face video generation model".

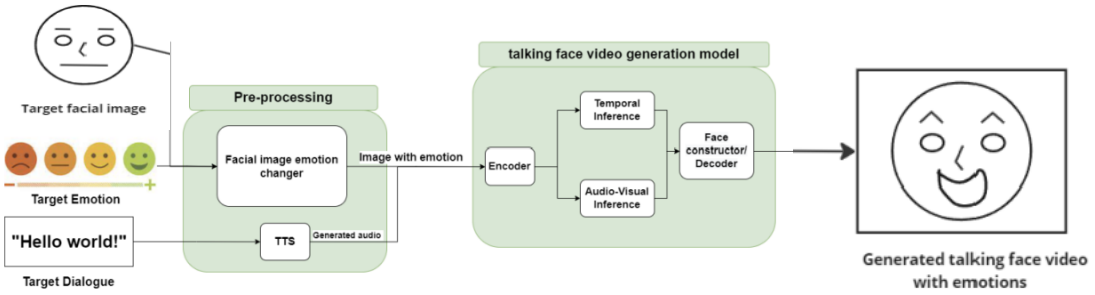

Fig.2. Second pipeline

A target facial image and the desired emotion are input into an emotion generator. This component generates a new modified image with the target emotion. Subsequently, the video generation model takes this emotionally modified image and incorporates lip synchronisation to create the final emotional talking face video.

Building upon the work of last year's students who developed the second pipeline, our objective is to enhance and refine their project. By employing this pipeline strategy, the need arises to discover or develop novel models that can effectively alter the emotion of the input image, contributing to the overall success of the emotional talking face video generation.

3. Emotion Generator

Generative adversarial networks (GANs) consist of two main components: a discriminator and a generator. The discriminator learns to differentiate between real and fake samples, while the generator aims to produce fake samples that closely resemble real ones. In the previous year, students explored GANs through two models: one using StyleGAN [3] with pre-trained weights and another training a new CycleGAN [4] model to transfer emotions. Despite their efforts, both StyleGAN [3] and CycleGAN [4] produced unrealistic images, leading to unsuccessful lip synchronisation.

Given these challenges, we decided to focus on Conditional GANs, a GAN-based approach for conditional image generation. In prior studies, Conditional GANs have been successfully used for tasks like domain transfer, super-resolution imaging, and photo editing by providing both the discriminator and generator with class information as a condition for generating samples. To tackle this, we adopted the scalable GAN framework called StarGAN [5], which allows flexible image translation across various target domains by providing conditional domain information. StarGAN's primary objective is to train a single generator (G) capable of learning mappings between multiple domains. The framework comprises a discriminator that distinguishes real from fake images and classifies real images into their corresponding domains. Meanwhile, the generator takes both the input image and the target domain label, spatially replicates and concatenates the target domain label with the image, and generates a fake image. During training, the generator aims to reconstruct the original image from the fake one, given the original domain label, producing images that are indistinguishable from real ones and can be classified as the target domain by the discriminator.

For training the model, we initially planned to utilise the RaFD [6] dataset, which provides eight labels for facial expressions, such as 'happy,' 'angry,' and 'sad.' However, while awaiting approval for the RaFD [6] dataset, we used the RAF [7] dataset, which contains seven labels for facial expressions, and the AffectNet [8] dataset, which offers eight labels for facial expressions, to train StarGAN [5]. Currently, we are actively working on improving the emotion generator by exploring better loss functions. The images provided below showcase the best results achieved thus far in our progress.

<Anne - Happy> <Anne - Input> <Anne - Disgusted>

Fig.3. Emotion results with Anne’s face

4. Lip synchronisation generator

The second stage of our pipeline involves lip synchronisation. In this stage, we use a single image with added emotion and audio input to generate a talking face.

We first decided to work with Wav2Lip [9], as it is a well-known open-source library. Our testing of Wav2lip on our images was able to perform an accurate lip synchronisation, however, it resulted in blurry mouth region and loss of teeth details. Last year's students used Wav2Lip and added a Spatial Transformer Network (STN), which performs different spatial transformations including rotation and translation. This was applied to faces to adjust the angle of the face and make them upright and centred. Then, they used a deepface lab to improve the quality of images.

We considered developing a network that could predict missing teeth regions and improve the quality of the mouth. To implement this, we tried with the Codebook Lookup Transformer (Codeformer) [10]. Codeformer is designed to learn from high-quality segments of the images or videos and use this learning to predict and enhance lower-quality areas. However, this approach still produced unrealistic teeth, overly smoothed mouth regions, and unstable facial appearance.

Another technique we explored was called video-retalking [11], which resulted in successful lip synchronisation without any loss in the teeth region. Video-retalking neutralises the mouth shape, regardless of the original expression (happy, sad, etc.), before proceeding with applying lip synchronisation [11]. For our project goals, we found this to be a limitation as we aim to preserve emotion in the image and deliver facial expressions while talking. As a result, we are currently in the stage of looking into another technique called SadTalker [12] and are considering this as an option that we could further improve.

4.1. Results shown with images

Fig.4. Lip synchronisation of different methods

4.2. Evaluation metrics for lip synchronisation

To evaluate the lip synchronisation aspect, we have so far used metrics such as MRE (Mean Relative Error), PSNR (Peak Signal-to-Noise Ratio), SSIM (Structural Similarity Index Measure)[13], and FID (Frechet Inception Distance) metrics. These are used to assess different aspects of video quality

including elements like noise ratios and structural similarity. Moreover, we are in the process of looking into LSE (introduced by Wav2Lip developer) [9] and FFHQ metrics as well.

	PSNR ⇧	MSE ⇩	SSIM ⇧	FID ⇩
Wav2lip	36.236	46.699	0.963	17.459
Last year's work	33.791	82.680	0.945	19.609
PC-AVS (without head pose)	14.985	6197.761	0.453	222.123
Video-retalking	34.958	62.498	0.955	74.156
Wav2lip+Codeformer	33.571	86.141	0.958	58.836
Videoretalking+Codeformer	32.552	108.669	0.951	26.042

Fig.5. Evaluation with ground truth video

* For SadTalker we will update the evaluation metrics once it is done as we are still working on it.

We are currently working on understanding why Wav2Lip receives the highest scores, despite its visually poor results.
Apart from Wav2Lip, our evaluation metrics show that video-retalking performs the best in video quality. However, as we identified a potential limitation with this approach, we are in the process of exploring the integration of SadTalker lip synchronisation and quality improvement.

5. Head pose implementation

We attempted to apply our images to the Pose-controllable talking face generation (PC-AVS) [14] model, which generates a head pose based on a reference video. Unfortunately, this caused the eye and lip region from the reference video to be replicated, such as makeup, and also resulted in face distortion. To overcome this limitation, we tried to apply a median filter to blur details in the eye and lip regions. This involves replacing each pixel's value with the median value of its neighbouring pixels to reduce the variance in colour. However, we could not see significant improvements.
We also explored another technology for the head pose, called the Thin Plate Spline Motion Model [15]. This model performs well when the face is facing forward, but it appears slightly unnatural when the face is not facing forward. We are looking into this technique to make improvements.

6. Future plans

There are several key areas for the improvement of our project. For the emotion generator, we are focusing on the improvement of facial expression generation particularly on sentiment expressions. Furthermore, for lip synchronisation, we are in the process of enhancing the quality of the face and the video. When lip synchronisation generation is successfully completed, we aim to incorporate lip synchronisation and head movement together to generate more realistic and natural movements for both the lips and the head. To ensure our system performs as expected, we are planning to implement additional evaluation metrics. Finally, we aim to incorporate Text-to-Speech (TTS) into our system and put all other stages together for the final demonstration.

Reference

[1] S. E. Eskimez, Y. Zhang, and Z. Duan, “Speech Driven Talking Face Generation from a Single Image and an Emotion Condition,” Aug. 2020, [Online]. Available: http://arxiv.org/abs/2008.03592

[2] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, “CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset,” IEEE Trans Affect Comput, vol. 5, no.4, pp. 377–390, 2014, doi: 10.1109/TAFFC.2014.2336244.

[3] T. Karras, S. Laine, and T. Aila, “A Style-Based Generator Architecture for Generative Adversarial Networks,” Dec. 2018, doi: 10.48550/arxiv.1812.04948.

[4] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks,” Mar. 2017, doi: 10.48550/arxiv.1703.10593.

[5] C.Yunjey, C.Minje, K.Munyoung, H.Jung-Woo, K.Sunghun and C.Jaegul Choo, “StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation,” Sep. 2018, doi: https://arxiv.org/abs/1711.09020

[6] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. Van Knippenberg. Presentation and validation of the radboud faces database. Cognition and Emotion, 24(8):1377–1388, 2010. 2

[7] L.Shan, D.Weihong and D.JunPing “Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild,” 2018, doi:http://www.whdeng.cn/raf/model1.html#dataset

[8] M.Ali, H.Behzad, Mohammad H. Mahoor. “AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild,” Oct.2017, doi:https://arxiv.org/abs/1708.03985

[9] K R Prajwal, Rudrabha Mukhopadhyay, Vinay Namboodiri, C V Jawahar, “A Lip Sync Expert Is All You Need for Speech to Lip Generation In The Wild”, Aug 2020, Available: https://arxiv.org/abs/2008.10010

[10] Shangchen Zhou, Kelvin C.K. Chan, Chongyi Li, Chen Change Loy, “Towards Robust Blind Face Restoration with Codebook Lookup Transformer”, Nov 2022, Available: https://arxiv.org/abs/2206.11253

[11] Kun Cheng, Xiaodong Cun, Yong Zhang, Menghan Xia, Fei Yin, Mingrui Zhu, Xuan Wang, Jue Wang, Nannan Wang, “VideoReTalking: Audio-based Lip Synchronization for Talking Head Video Editing In the Wild”, Nov 2022, Available: https://arxiv.org/abs/2211.14758

[12] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, Fei Wang, “SadTalker: Learning Realistic 3D Motion Coefficients for Stylized Audio-Driven Single Image Talking Face Animation”, Nov 2022, Available: https://arxiv.org/abs/2211.12194

[13] Umme Sara, Morium Akter, Mohammad Shorif Uddin, “ Image Quality Assessment through FSIM, SSIM, MSE and PSNR”, Journal of Computer and Communications, March 2019, DOI:10.4236/jcc.2019.73002

[14] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, Ziwei Liu, “Pose-Controllable Talking Face Generation by Implicitly Modularized Audio-Visual Representation”, Apr 2021, Available: https://arxiv.org/abs/2104.11116

[15] Jian Zhao Hui Zhang, “Thin-Plate Spline Motion Model for Image Animation”, Mar 2022, Available: https://arxiv.org/abs/2203.14367