Department of Electrical, Computer, and Software Engineering

Part IV Research Project

Literature Review and

Statement of Research Intent

Project Number: 64

Talking face generation System

Yugyeong Hong

Gayeon Kim

Ho Seok An, Trevor Gee

26/04/2023

Declaration of Originality

This report is my own unaided work and was not copied from nor written in collaboration with any other person.

Name: Yugyeong Hong

ABSTRACT: Significant progress has been made in the field of talking face generation, but existing methods often neglect facial emotion or rely on audio or video input. This paper aims to explore existing projects and advancements in emotional talking face generation. This research will provide insights to improve existing projects, generate new ideas for solutions, and deepen our understanding of previous work, enabling us to anticipate and tackle challenges throughout the year to complete the project.

1. Introduction

A talking face generation system can be utilized for multiple purposes, including animation, virtual assistant development, enhancement of chatbot interaction, and in healthcare field. Although significant advancements have been achieved in this field, a majority of the research has focused on generating realistic synchronization, identity preservation eye blinks or head motion in the synthesized talking face video [1][2][3][4]. Emotion is a vital component of communication, as conveyed messages are not just limited to spoken words, but also involve tone and nonverbal emotional cues. Emotion has a direct impact on the meaning of the transmitted message, and it can significantly alter its interpretation [5]. Emotional talking face generation systems offer a novel and exciting tool for a wide range of applications, ranging from general entertainment to healthcare. One such example is the development of interactive chatbots with emotional capabilities, which would make human-computer interactions more natural and meaningful. Another potential use is in healthcare, where the technology could be used to teach emotion recognition to individuals who struggle to understand emotions in others, without the need for uncomfortable human interaction [6]. Additionally, it could improve accessibility for people who are hard of hearing and rely on visual cues. By generating visual emotions, messages could be delivered more clearly and with more nuance, as emotional intent would be more apparent. This is especially important given that humans have difficulty recognizing emotions based solely on speech [7]. This paper will conduct a literature survey to examine existing methods for emotional face generation and how emotions are handled in these systems. The insights gained from this analysis will inform the research goals and objectives.

2. Literature Review

In this literature review, I have examined the text-driven emotional talking face generation developed by Alan Lin and Takahiro Ishiguro in 2022, which includes temporal project issues, as well as other methods for generating emotional talking faces.

2.1 Temporal development

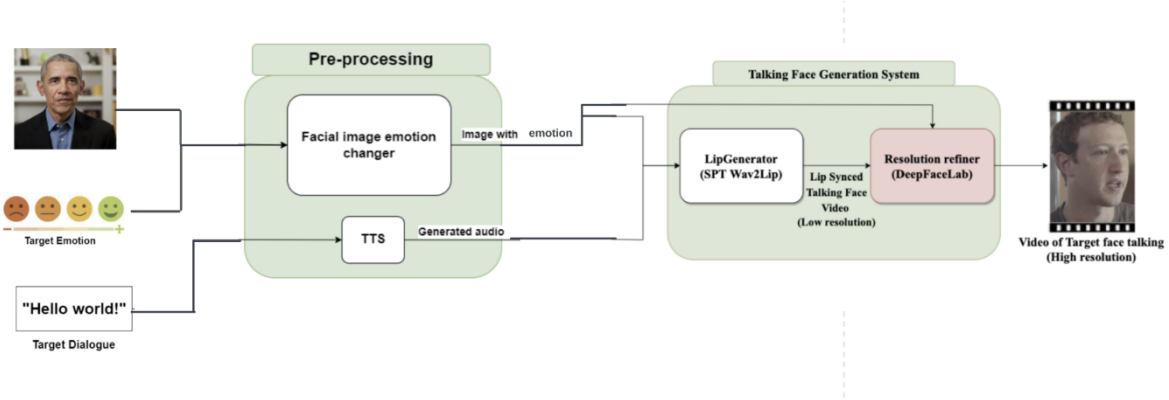

Fig 1. Final pipeline of Talking face generation system

The approach taken by Alan and Takahiro in emotional talking face generation is to split up the generation of emotional expressions and lip sync into two distinct steps, each with its own dedicated model, instead of generating both in a single model. By taking a single facial image, a categorical emotion label, and text as inputs, this approach can generate a synchronized talking face video that expresses the conditioned emotion.

During the pre-processing stage, a target facial image and target emotion are passed into a “facial image emotion changer”. The facial image emotion changer of the input image is accomplished through a StyleGan [8] implementation using pre-trained NVIDIA weights, which allows for the transfer of image emotions and emotional latent direction. To change the emotion of the input image, it is first encoded into a latent space, creating a latent representation of the image. This latent representation is then modified using the pre-trained emotional direction, which produces a new modified image with the target emotion. Generating edited images from modified latent codes is the main function of StyleGan [8]. The resulting modified image is then fed into the talking face video generation model, which adds lip synchronization to the image to generate a final talking face video that effectively communicates the intended emotion through realistic facial expressions and lip movements that match the spoken words.

During the talking face video generation model stage, a target facial image with changed emotion and generated audio from text-to-speech (TTS) are passed into a “LipGenerator”. The LipGenerator was made to adopt Wav2Lip [10] which uses a simple encoder-decoder framework, making the whole generation system a black box and focusing more on providing customized input and feedback. To develop Wav2Lip [10], Alan and Takahiro used the VoxCeleb2 [11] dataset, which consists of video recordings of celebrity interviews that were extracted from YouTube. This dataset was preferred over the LRS2 [12] dataset, as it contains significantly more videos of spoken utterances. The SPT-Wav2Lip model, developed by Alan and Takahiro, represents an enhanced version of the Wav2Lip [10] model that incorporates a

spatial transformer network [13]. This network addresses the issue of large datasets by incorporating a mechanism within the neural network that can generalize the head-poses observed. The spatial transformer network is placed in front of the facial encoder network, allowing facial transformations to be performed before encoding. To enhance the visual quality of their results, Alan and Takahiro leveraged the capabilities of deep-fake technology, which enables the transfer of face images between different sources. By utilizing high-quality video inputs, they were able to extract the faces from each individual frame and use them as a replacement for images that were noisy or blurry. This approach was implemented using an open-source deep-fake software, DeepFaceLab [9].

Therefore, the final pipeline shown in Fig 1 shows that the emotional facial image created on the pre-processing stage generates emotional talking face video through LipGenerator with audio file generated by TTS.

2.1.1 Challenges of temporal development

Although a new model called SPT-Wav2Lip was created and Deep-fake technology was used to improve the quality of generated talking face videos, the system still faces difficulty in generating accurate lip-synthesized frames when a speaker pauses their speech. For example, when the speaker's mouth should be closed at the end of a sentence, the system may produce a frame where the lips continue to move. While using DeepFaceLab [9] can address the visual quality issue, it is only effective when multiple images of the target speaker are provided and not when only a single image is available. Since the project requires a single image as input, an alternative solution should be explored to replace the DeepFaceLab [9] approach. In addition, the current state of image-to-video talking face technology is still far from realistic. One of the main reasons for this is that the generated videos lack head movement, resulting in a static image with only lip movement. Furthermore, replicating emotions using StyleGan's [8] latent spaces is challenging as it struggles to differentiate between negative emotions and neutral ones. In addition, I have observed that our model based on Wav2Lip [10] tends to generate frames where the inside of the mouth is filled with black in most of the frames, leading to the loss of details such as teeth. Fig 2 illustrates a sample of the generated frame in temporal project model.

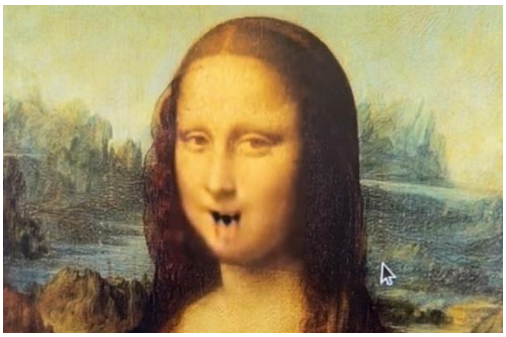

Fig 2. Generated image in temporal project model.

2.2 Emotional Talking Face Generation

I have come across three distinct methods for generating emotional talking faces. The first approach is a speech-feature-driven system that directly maps speech features to video frames. The second approach is a two-stage process that involves converting the speech input to face landmarks, followed by estimating video frames using the predicted landmarks. The third approach is an end-to-end system that generates talking faces directly from a conditioned image and the speech signal. Note that our project implemented an end-to-end system.

2.2.1 Speech-feature-driven method

[14] generates emotional talking face videos by utilizing a 3DMM, a technique used for 3D face reconstruction, to capture significant facial movements and a StyleGan-based texture map to capture fine details and subtle variations in appearance. The emotional information is present in both the 3DMM and texture map, and these can be modified by neural networks in a continuous manner. Additionally, the smoothing process can be easily achieved by averaging in the coefficient/latent spaces. To improve their model's fidelity, the authors of [14] incorporated teeth filling module and textures generated by StyleGan, allowing for high-quality outputs at a resolution of up to 1024*1024. This approach [14] could be useful for enhancing our current project, but it should be noted that it relies on video input to produce emotional talking face videos.

Experiments conducted by [14] used two datasets: MEAD [15] and RAVDESS [16]. The MEAD dataset [15] comprises high-quality talking-face videos of 60 actors expressing eight categories of emotions at three intensity levels, recorded from seven different view angles. Each actor has around 30 videos for each intensity of each emotion at each view angle. To demonstrate the model's generalizability, the RAVDESS dataset [16] was used for the experiment.

2.2.2 Two-stage method

The Emotion-Aware Motion Model (EAMM) [17] is a method that generates emotional talking faces in a one-shot manner by utilizing an emotion source video. [17] is comprised of two main components: an Audio2Facial-Dynamics module (A2FD) that generates audio-driven talking faces with neutral expressions from a single neutral frame, and an Implicit Emotion Displacement Learner for extracting emotional pattern that represents emotion-related facial dynamics as linearly additive displacements to the previously acquired motion representation. In the A2FD module, the model employs three encoders to extract relevant information from three inputs, namely source image feature, audio feature, and pose feature. The three extracted features are then combined and fed into a LSTM-based decoder to recurrently predict unsupervised motion representations for the entire sequence. The implementation of a key-point detector and the adoption of a flow estimator to generate a dense warping field with the Implicit Emotion Displacement Learner module

allow for the generation of an emotional talking face video by adding extracted emotion pattern onto the motion representations of an arbitrary person. The researchers in [17] noted that although emotional information can be retrained, there are several undesirable distortions present around the face boundary and the mouth. This is because the computed displacements include not only emotion features but also other aspects, such as identity, pose and speech content, which results in inaccurate guidance for the subsequent generation.

The A2FD module of the [17] model is trained on the LRW dataset [18], which does not have any emotion annotation. LRW [18] is a collection of in-the-wild audio-visual data from BBC news, comprising 1000 utterances of 500 distinct words, with each lasting about one second. On the other hand, the Implicit Emotion Displacement Learner module is trained on the emotional dataset MEAD [15].

2.2.3 end-to-end method

Our temporal project relies on StyleGan [8] for facial expression editing and Wav2Lip [10] for lip synchronisation. Generating edited images from modified latent codes is the main function of StyleGan [8]. However, identifying decoupled editing directions for different facial expressions is a challenging and time-consuming task because these expressions are often correlated with other attributes such as poses in the latent space of StyleGan, which is caused by biases in the training data [19]. There are some methods directly designed for facial expression editing of image. ExprGAN [20] is a type of conditional-GAN method that could convert face images into specific expressions with a continuous range of intensities, even when the intensity labels are absent from the training data. To achieve this, an expression controller module is utilized that introduces uniform noise to the initial one-hot label. StyleRig [21] applies 3DMM, a technique used for 3D face reconstruction, in the StyleGan-based editing process. This involves mapping the 3DMM coefficients to the latent code of StyleGan, providing explicit control over the expression and pose of the generated faces. The method proposed in [22] generates an edited image by first reconstructing a 3D Morphable Model (3DMM) from the input image. Then, the shape and texture are modified in two separate branches based on the input 3DMM coefficients. Consequently, these methods can be applied to improve our current project enhancing emotional image quality.

2.2.4 Comparing different methods.

Table 1 presents the strengths and weaknesses of each method for generating emotional talking faces. There are different techniques that can be utilized in each method to implement the generation of talking faces with emotion. For instance, the two-stage method can use either a 2D or 3D face landmarks generator or a neural network to generate face landmarks.

In addition, the two-stage and the end-to-end approach can be applied with various lip synchronization methods. Therefore, firstly we should consider which method is more appropriate to improve our current project.

Table 1: Comparing different methods in handling emotional talking face generation.

	Advantages	Disadvantages	Reference
Speech-feature-driven	-Emotion displayed across whole face	- Emotion generated relies on input video. -Rely on long video recording of a source portrait	[14]
Two-stage	-Potentially more control of specific emotion -High potential for expression i.e., emotion in head poses	- Some approaches have low range of emotions - May have undesirable artifacts around face boundary and the mouth	[17] [23]
End-to-end	-Control of specific emotion	-Degrade image or video quality -Limited by the speech emotion recognition accuracy	Our project

3. Evaluation Metric

The evaluation of talking face generation models has become more standardized and qualitative measurements are now used. Several metrics, such as PSNR, SSIM, MSE, and VMAF, can be used to measure the quality of generated frames, while metrics such as FID and CPBD can evaluate the overall quality of a generated video. M-LMD can evaluate in two-stage method to measure landmarks distances on the mouth. F-LMD to measure the accuracy of facial expression and poses can be used as well. In addition, LSE-D and LSE-C can be used to assess the lip synchronization of generated video.

4. Research Intent

Based on my thorough literature review, I believe that improving our emotional generator and enhancing the video quality in the temporal project can improve the quality of the generated videos without replacing the lip synchronization model. For instance, generating a better emotional image by utilizing techniques such as ExprGAN [20] and StyleRig [21], as explained in section 2.2.3 can lead to improved quality of emotional talking face video. The fundamental issue of Wav2Lip [10] model which generate frames where the inside of the mouth is filled with black, resulting in the loss of details such as teeth can be solved by implementing teeth filling module [14]. In addition, training Wav2Lip [10] at higher resolutions and using super-resolution techniques on the SPT-Wav2Lip model can replace DeepFaceLab [9], leading to an enhanced video quality. Hence, combining the techniques outlined above appears to enhance the overall

quality of emotional talking face generation. Nonetheless, we will consider substituting Wav2Lip's generator model with a more recent trending model such as the diffusion model as it may be more appropriate. Additionally, we will explore the two-stage method that involves implementing facial landmarks or extracting emotional features and adding them to the lip-synchronized video. This is because the two-stage approach has the potential to provide more control of specific emotion and head pose. It is also crucial to experiment with new datasets such as MEAD [15], as our project is based on the VoxCeleb2 [11] dataset, which was not initially designed for emotional video generation. Ultimately, we aim to generate a realistic emotional talking face video and text input, with the following objectives:

- Create an upgraded emotional image using different approaches such as ExprGAN [20] and StyleRig [21].
- Replace DeepFaceLab [9] by training Wav2Lip [10] at higher resolutions and using super-resolution techniques.
- Replace Wav2Lip generator model.
- Create a model based on the Two-stage approach implementing feature extractor.
- Implement head pose using landmark approach or pre-defined pose.

After conducting a literature review, we plan to apply our gathered knowledge to develop and test our proposed idea. Our goal is to create a realistic talking face generation model by the end of the mid-semester one break. Initially, we will work on improving the current talking emotional face generation developed by Alan and Takihiro by replacing the emotional changer in the pre-processing stage and utilizing DeepFaceLab technology. We will also explore the two-stage method, which involves implementing head pose and realistic emotion generation using either 3D landmark technology or encoder-decoder based technology. In Fig 3, it represents the simple pipeline of the two-stage model we aim to create. Our aim is to complete the model development by the end of the semester break, with the remaining time devoted to perfecting the documentation and presentation.

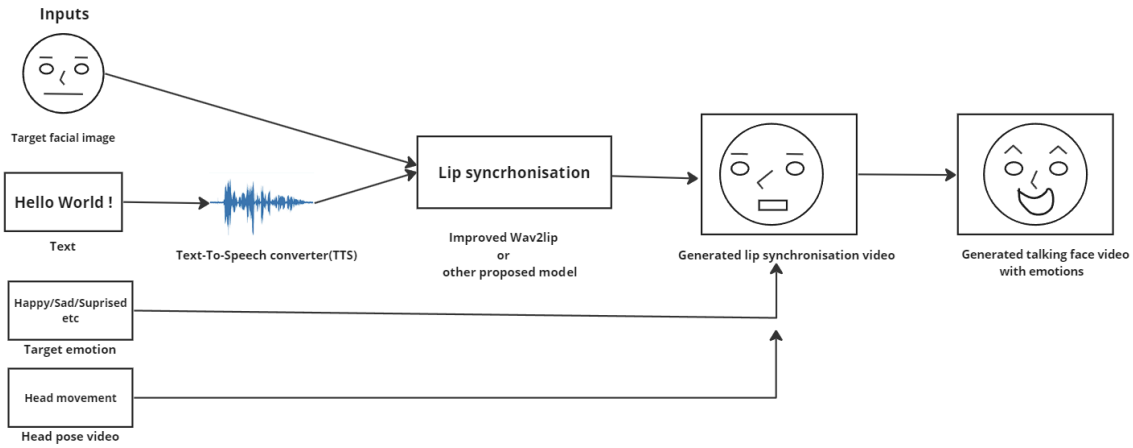

Fig 3. Simple pipeline we designed to develop.

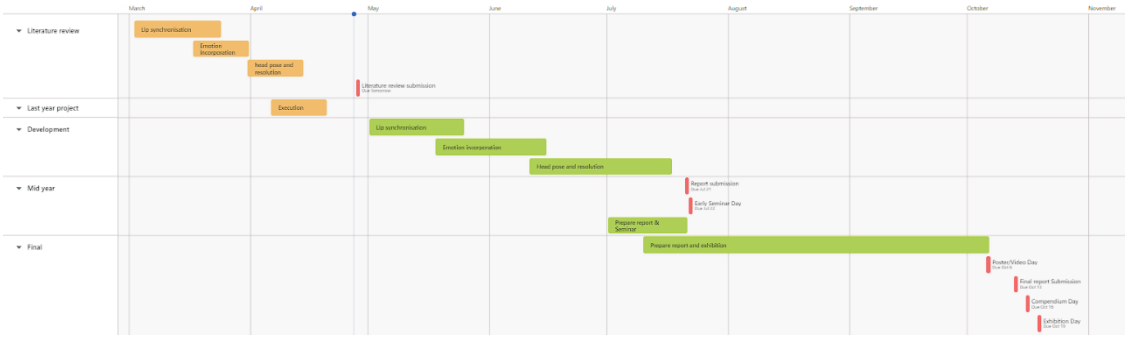

Fig 4. Our Gantt chart for project

5. Conclusions

In conclusion, the development of accurate and realistic emotional talking face videos is still a challenging task, and there are various approaches to generate such videos, including speech-feature-driven methods, two-stage methods, and end-to-end methods. The end-to-end method, which involves generating talking faces directly from a conditioned image and speech signal, is the approach used in the temporal project discussed in the text. The project relies on StyleGan for facial expression editing and Wav2Lip for lip synchronization. However, there are still limitations in the technology, such as difficulty in generating accurate lip-synthesized frames when a speaker pauses their speech and replicating emotions using StyleGan's latent spaces. Overall, the development of emotional talking face videos requires further research and improvement to achieve more realistic and accurate results. This paper presents the approaches undertaken by previous researchers in improving the existing emotional talking face generation systems. The literature review has extensively examined numerous techniques aimed at generating high-resolution videos of emotional talking faces that appear realistic in terms of emotional expression.

References

[1] Z. Sibo, Y.Jiahong, L.Miao, Z.Liangjun, "Text2Video: Text-Driven Talking-Head Video Synthesis with Personalized Phoneme-Pose Dictionary," Jan.2022, [Online]. Available: https://arxiv.org/abs/2104.14631

[2] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, and D. Li, "MakeItTalk: Speaker-Aware Talking-Head Animation," Feb. 2021, [Online]. Available: https://arxiv.org/abs/2004.12992

[3] Z.Hang, S.Yasheng, W.Wayne, L.Chen, W.Xiaogang, L.Ziwei Liu, "Pose-Controllable Talking Face Generation by Implicitly ModularizedAudio-VisualRepresentation,"Apr.2021,[Online].Available:https://arxiv.org/abs/2104.11116

[4] W.Suzhen, L.Lincheng, D.Yu, Y.Xin, "One-shot Talking Face Generation from Single-speaker Audio-Visual Correlation Learning," Dec.2021, [Online]. Available: https://arxiv.org/abs/2112.02749

[5] M. Alpert, R. L. Kurtzberg, and A. J. Friedhoff, "Transient voice changes associated with emotional stimuli," Arch. Gen. Psychiatry, vol. 8, no. 4, pp. 362–365, 1963. doi: 10.1001/archpsyc.1963.01720100052006

[6] S. A. Cassidy et al., "Expressive visual text-to-speech as an assistive technology for individuals with autism spectrum conditions," Comput Vis Image Underst, vol. 148, pp. 193–200, Jul. 2016, doi: 10.1016/j.cviu.2015.08.011

[7] S. Eskimez, K. Imade, N. Yang, M. Sturge-Apple, Z. Duan, and W. Heinzelman, "Emotion classification: How does an automated system compare to Naive human coders?" Apr. 2016, pp. 2274–2278. doi:10.1109/ICASSP.2016.7472082.

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

[9] I. Perov et al, "DeepFaceLab: Integrated, flexible and extensible face-swapping framework," 2020. [Online]. Available: arXiv:2005.05535.

[10] K. R. Prajwal et al, "A Lip Sync Expert Is All You Need for Speech to Lip Generation in The Wild," 2020. [Online]. Available: arXiv:2008.10010.

[11] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," Jun. 2018, doi: 10.21437/interspeech.2018-1929.

[12] T. Afouras, J. S. Chung, A Senior and A. Zisserman, "Deep Audio-Visual Speech Recognition," 2018. [Online]. Available: arXiv:1809.02108.

[13] M. Jaderberg, K. Simonyan, A. Zisserman and K. Kavukcuoglu, "Spatial Transformer Networks," 2015. [Online]. Available: arXiv:1506.02025.

[14] S.Zhiyao, W.Yu-Hui, L.Tian, S.Yanan, Z.Ziyang, W.Yaoyuan, L.Yong-Jin, "Continuously Controllable Facial Expression Editing in Talking Face Videos," Sep. 2022. [Online]. Available: https://arxiv.org/abs/2209.08289

[15] K. Wang, Q. Wu, L. Song, Z. Yang, W. Wu, C. Qian, R. He, Y. Qiao, and C. C. Loy, "MEAD: A large-scale audio-visual dataset for emotional talking-face generation," in ECCV (21), vol. 12366, 2020, pp. 700–717.

[16] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, 13 multimodal set of facial and vocal expressions in north American English," PLOS ONE, vol. 13, no. 5, pp. 1–35, 05 2018.

[17] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Wayne Wu, Feng Xu, Xun Cao," EAMM: One-Shot Emotional Talking Face via Audio-Based Emotion-Aware Motion Model," Sep. 2022. [Online]. Available: https://arxiv.org/abs/2205.15278eamm

[18] J. S. Chung, A. Zisserman "Lip Reading in the Wild," Asian Conference on Computer Vision, 2016

[19] E. Harkonen, A. Hertzmann, J. Lehtinen, and S. Paris, "Ganspace: Discovering interpretable GAN controls," 2020.

[20] H. Ding, K. Sricharan, and R. Chellappa, "Exprgan: Facial expression editing with controllable expression intensity," in AAAI, 2018, pp. 6781–6788.

[21] A. Tewari, M. Elgharib, G. Bharaj, F. Bernard, H. Seidel, P. Perez, M. Zollhofer, and C. Theobalt, "Stylerig: Rigging stylegan for 3d ¨control over portrait images," in CVPR, 2020, pp. 6141–6150.

[22] Z. Geng, C. Cao, and S. Tulyakov, "3d guided fine-grained facemanipulation," in CVPR, 2019, pp. 9821–9830.

[23] S.Sanjana, B.Sandika, Y.Ravindra, B.Brojeshwar, "Emotion-Controllable Generalized Talking Face Generation," May. 2022. [Online]. Available: https://arxiv.org/abs/2205.01155