Department of Electrical, Computer, and Software Engineering

Part IV Research Project

Compendium Report

Project Number: 64

Emotional Talking Face Generation System using DNN

Author: Gayeon Kim, Yugyeong Hong

Supervisor(s): Ho Seok Ahn, Trevor Gee

17/10/2023

Declaration of Originality

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

Gayeon Kim

Yugyeong Hong

Name: Gayeon Kim, Yugyeong Hong

ABSTRACT: Our project focuses on creating realistic artificial humanoid avatars from a single image and text, enhancing human-robot interaction. Unlike traditional 3D modelling, we use machine-learned image augmentations to create avatars exhibiting accurate lip motion, head movements, and emotional expressions. Our unique approach integrates emotion processing with lip synchronisation, using text as input, thereby providing a more engaging experience. Our system outperforms existing open-source models in both performance and realism, although there is potential for further refinement.

1. Introduction

In the modern digital era, creating interactive and realistic talking faces has become a topic of significant interest due to advancements in technology. This field's relevance is emphasised by the expected growth of the global Human-Machine interface market, projected to increase from USD 4.9 billion in 2022 to USD 7.3 billion by 2027 at a compound annual growth rate (CAGR) of 8.1% [1]. This statistic indicates a growing need for intuitive communication methods with technology.

Talking faces serve as a human-like channel for interaction, easing the adoption barrier often linked with new technologies. Talking faces can be used for various purposes, including human-computer interaction, animation, healthcare, and entertainment. Despite their rising demand, realistic talking face generation faces challenges like emotion rendering, head pose dynamics, reliance on extensive inputs, and hardware cost efficiency.

Current research such as [2][3][4][5] typically examines lip motion separately from emotional expressions and head movements, yet emotions significantly influence communication clarity. [6] highlights the essential nature of emotional engagement in improving the clarity of communication, demonstrating that the absence of emotional signals results in a more than 30% increase in errors

during human-robot interactions. In addition, existing models often fail to attain a level of high-quality realism, becoming apparent, especially in instances where the models have not been fine-tuned for a specific individual. Furthermore, common dependency on audio or video inputs, as seen in studies [7][8][9][10], limits universal accessibility due to storage and transmission constraints, making a single-image and text-input method more practical. To address these gaps, we aim to develop a taking face generation system that not only conveys a variety of emotions, including happiness, sadness, and anger, but also outperforms other models in delivering high-fidelity realism.

Our approach, which utilizes a single image and text as inputs, improves accessibility and practicality, particularly in situations where employing a video and audio file is impractical or impossible. Confronting these existing challenges, our system is designed to accept an emotional state, a single facial image, and desired dialogue in text format and focuses on head pose implementation and high-quality realism.

In the following sections of this paper, we provide a detailed explanation of our research journey and the results of our experiments. Section 2 presents our project journey, including a comprehensive analysis of related works which lay the groundwork and establish the context for our investigation and our research approaches. Our research approach starts with an overview of our original End-to-End strategy and transitions to our improved Step-by-Step approach, which is our final methodology. A detailed explanation of the Step-by-Step approach follows, including their individual stages and the metrics used for assessment. In section 3, the results of our experiments are described, and in section 4, we outline the steps for setting up the development environment. Section 5 of this paper details the events of the seminar and outlines our plans for the exhibition day. Section 6 provides a conclusion to our paper, and Section 7 details the files included with this compendium report.

2. Project journey

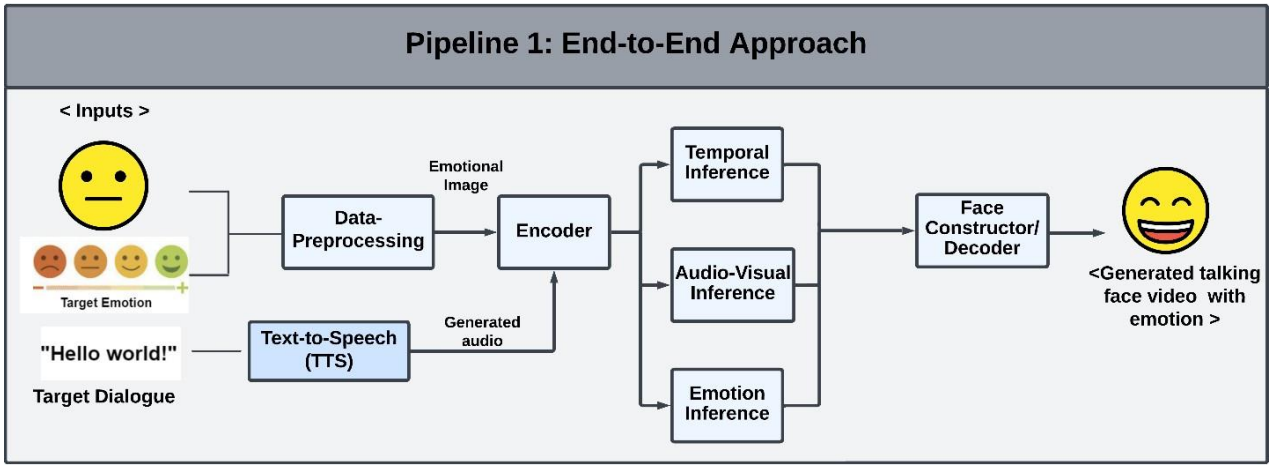
Before starting our project, a comprehensive literature review was conducted. Our research integrates various areas of image augmentation, such as modifications of facial expressions, lip movements, head positioning dynamics, and visual quality enhancements. In this section, we outline the key literature that supports our approach, also discussing how our implementation has evolved and been refined based on these studies.

Using a specified face, emotion, and dialogue (whether text or audio), the system creates a video where the chosen face conveys the dialogue with the desired emotion. Two different approaches exist for this purpose. The End-to-End approach combines emotion generation, lip synchronization and head movement within a single model. In contrast, the Step-by-Step approach divides these tasks into separate stages and models.

2.1.End-to-End

Initially, the aim was to modify emotion and synchronise lip and head movements using a single model. The comprehensive flow of this approach is illustrated in Fig. 1. Our strategy was developed based on the EmoTalkingFace, introduced in [11], that had pre-established capabilities for producing emotional faces. This involved inputting a facial image, emotion, and dialogue, which was converted to audio using text-to-speech. The unified model then produced a video with the face expressing the chosen emotion. However, due to lengthy training times and challenges in achieving synchronised lip movements, head motions, and video quality, a step-by-step method was adopted.

Fig. 1. End-to-End Approach pipeline



2.2.Step-by-Step

In contrast to the original pipeline that combined tasks like lip synchronisation, emotion generation, and head movements into a single model, the updated pipeline divides these functions into five distinct stages, as described in Fig.1, each powered by its own specialised model.

The process unfolds in five steps:

1. Emotional Face Image Generation: A target facial image and emotion are input to produce an emotional image.
2. Talking Face Video Generation with Lip Synchronisation: The emotional image is combined with generated audio (from the target dialogue converted via text-to-speech) to produce a video with synced lip movements.
3. Paste-back Operation: This stage pastes the cropped emotional image back to its original image.
4. Head Pose Implementation: Implementing head movements to make the video more dynamic and realistic.

5. Visual Quality Improvement: Enhancements are made to the video's visual quality, resulting in the final output: a talking face video with the desired emotion and head movements.

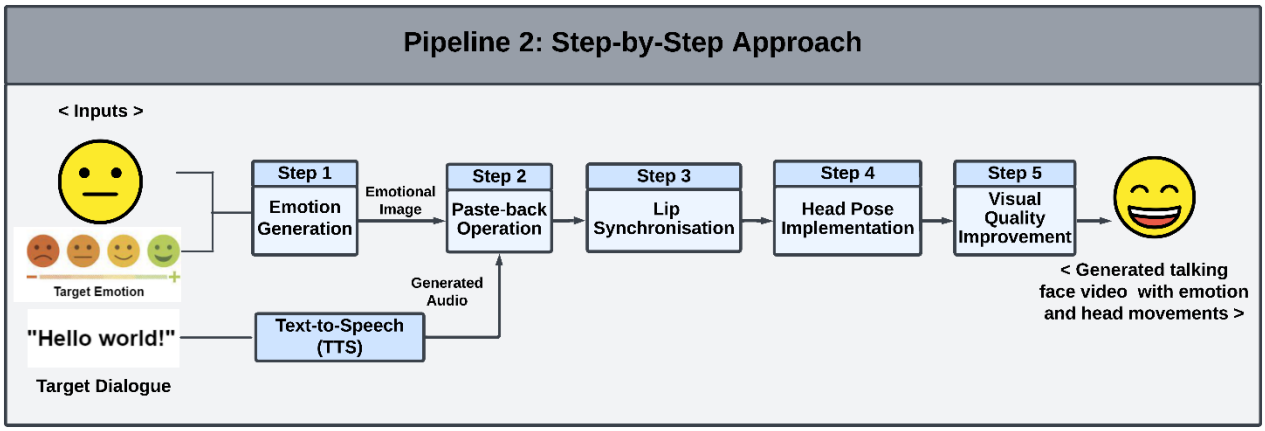


Fig. 2. Step-by-Step Approach pipeline

2.2.1. Emotion

Emotions play a significant role in how we communicate. Facial expressions act as indicators of feelings like happiness, sadness, or anger. Recognizing their importance in daily interactions, we worked on the Emotion Generation step, which focuses on emotion synthesis and improving visual quality. Initially, we investigated the capabilities of Generative Adversarial Networks (GANs) known for producing realistic images through a two-network setup: a generator that creates images and a discriminator that differentiates between generated and real images.

We started with Style-GAN [12], from the "style-transfer" domain, to produce emotional facial images. Unlike traditional GANs, Style-GAN changes its generator, using the progressive GAN approach and adjusting the latent space input. These advancements led to a new style editing method. However, StyleGan had difficulties distinguishing between neutral and negative emotions, as shown in Fig.3. This led us to explore Conditional GANs for better image creation.

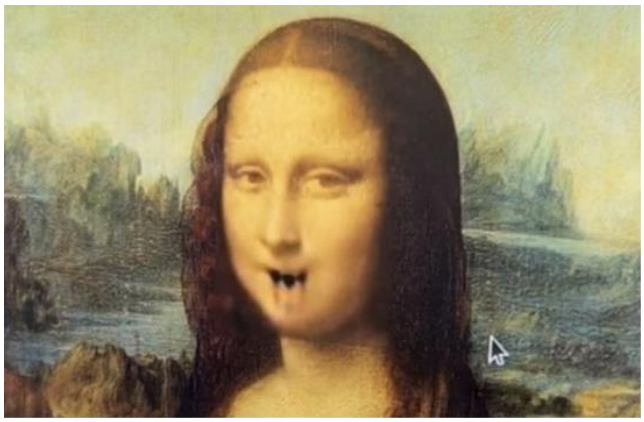


Fig.3. Style-GAN based emotion generator result

StarGAN [13], an advanced version of Conditional GANs, allows for flexible image translation across various domains using conditional domain data. It uses a single generator to map domain relationships and a discriminator to sort through real and created images, categorizing the real ones into different domains. Regarding our training datasets, while we aimed for RaFD [14], we initially used RAF [15] and AffectNet [16] due to approval delays. The training goal was to ensure accurate recreation from generated images and alignment with the intended domain. However, StarGAN sometimes produced unclear emotional images, evident in Fig. 4.

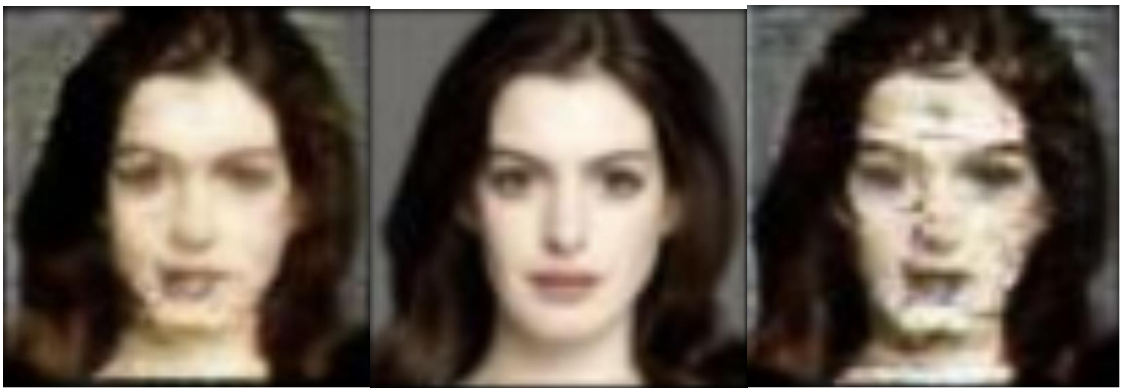


Fig.4. StarGAN approach (a) Happy, (b) Neutral, (c) Angry

Our next approach for emotional face generation was using audio-visual datasets, a field not yet widely explored due to limited labelled emotional audio-visual datasets. Our focus was on The

Emotion-Aware Motion Model (EAMM) [17]. This model includes the Audio2Facial-Dynamics (A2FD) module for neutral talking faces and the Implicit Emotion Displacement Learner to detect emotion-driven facial changes. Despite its approach, it had challenges, especially when blending unnecessary features, leading to facial distortions. Our assessments pointed out certain issues in its emotional accuracy and visual quality, shown in Fig. 5.

Fig.5. EAMM approach (Happy)

In the final stages of our research, we explored the advanced capabilities of EmoTalkingFace [18]. This model integrates an emotional discriminator designed to recognize emotions from video data. Adding LSTM layers to the generator improved the generation and transition of facial emotions across video sequences. This combination of video emotion classifiers with generative models showed potential for more accurate emotional expression generation.

The design of the emotional discriminator was notable. It added an additional class for fake videos based on BAGAN and included five layers of 2-D convolutional layer, followed by two FC layers. This discriminator processes each video frame and directs the resulting sequence through an LSTM layer. The final output from the LSTM is then sent to an FC layer, producing the probabilities of

seven classes, six of which represent emotions based on Ekman’s emotion model, with the seventh being a 'fake' category. For training, we used the CREMA-D dataset, which is based on Ekman’s emotion model. We calculated the categorical cross-entropy loss by comparing the emotion labels of real videos against the 'fake' label for generated outputs.

A challenge was the visual distortions in the generated emotional images. To address this, we integrated Codeformer [19], a tool that focuses on improving visual quality and ensuring that generated facial features align with the intended emotions. Combining this with EmoTalkingFace's framework, we achieved a system that produces images with improved clarity and emotional accuracy.

Shifting our focus from producing videos to generating high-quality emotional images, the results were positive. The produced images not only had better clarity but also captured the intended emotional expressions, as shown in Fig. 6. Our journey in this research highlights the importance of continuous improvement and adaptability in the ever-changing field of image generation.

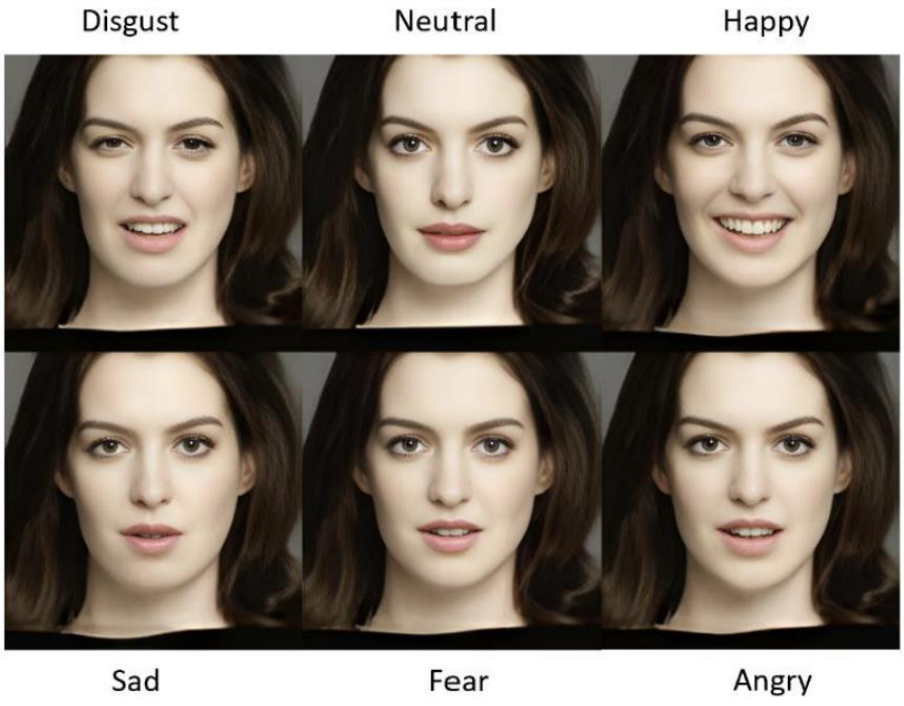


Fig.6. Final approach with visual quality enhancement

2.2.2. Lip synchronisation

Lip synchronization is critical in creating realistic talking faces, enhancing speech comprehension through visual cues. Various techniques have emerged from significant research in this field, particularly focusing on Generative Adversarial Networks (GANs) for their ability to produce realistic images. GAN consists of two neural networks: the generator, which generates a single image, and the discriminator, which distinguishes the image produced by the generator from real data. For instance, the LipGAN [30] model employs GANs to synchronize lip movements with corresponding audio converted to melspectograms by predicting lip shapes for sequential video frames. However, its reliance on a pixel-wise loss function leads to general predictions rather than precise lip-audio correlations. To address this, the Wav2Lip [20] model was introduced, prioritising accurate mouth shapes through discriminative training but falling short by generating low-resolution, blurry outputs. An attempt to integrate GANs with Wav2Lip showed minor improvements in synchronisation without achieving the desired quality.

A more advanced model, Sadtalker [21], builds on Wav2Lip's foundations, using the 3D Morphable Model (3DMM) technique for detailed facial reconstructions, thereby enhancing the realism of lip movements and overall facial expressions.

Therefore, in our lip synchronisation stage, we decided to utilise pre-trained models and weights from Sadtalker. Sadtalker’s unique lip motion coefficients serve as the target within the established Wav2lip network. Then, a 3D Morphable Model (3DMM) technique is employed to reconstruct the 3D facial structure, with a focus on lip movements, enabling precise lip synchronisation. We focus solely on the lips and disregard other facial movements for our project. Fig. 7. Shows our lip synchronisation results.

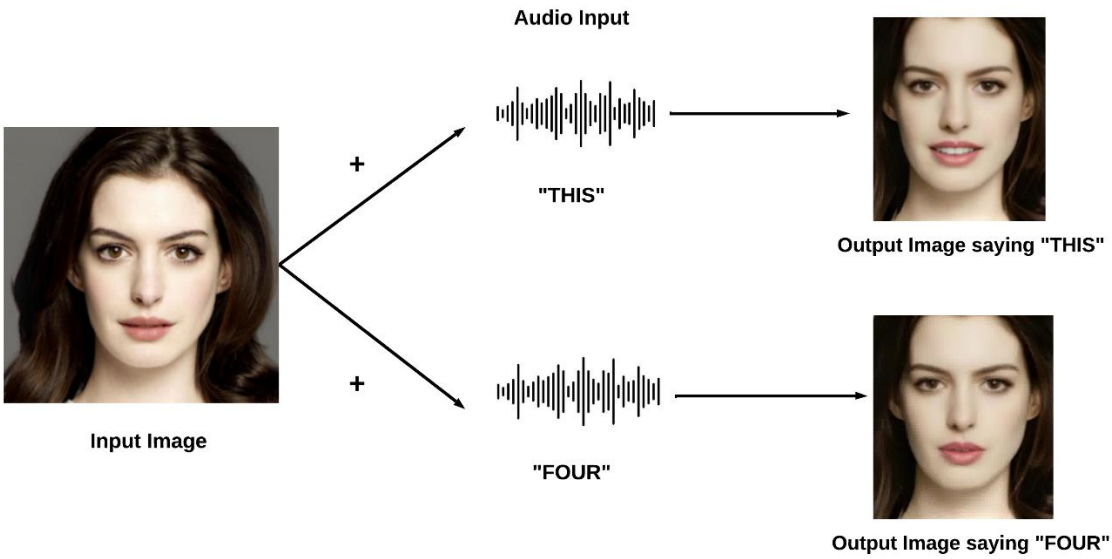


Fig.7. Lip synchronisation results

2.2.3. Paste-back operation

In our process, adding head pose movement after lip synchronization presented a challenge due to the initial image cropping focused solely on the face, omitting key spatial references like the neck and shoulders essential for natural head movement. To resolve this, we implemented a "paste-back" operation before the head pose stage, where the emotion-processed face is reinserted into the original, uncropped image. This step is implemented before lip synchronization as it is applied at the image level, while lip synchronization involves video processing.

In the first attempt at the paste-back operation, Haarcascades face detection method was used to crop and overlay the central facial region onto the original image, deliberately excluding the forehead and chin. Despite applying Gaussian blur for smoother boundaries, the integration appeared unnatural, with clear edges noticeable, as shown in Fig. 8(a). Complications also arose due to size discrepancies between the cropped and original faces.

In the second approach, the focus shifted to essential facial features—eyebrows, eyes, nose, and mouth—due to their critical role in conveying emotion. The same detection and landmarking methods were employed, but only these key features were extracted. This strategy, while successful in aligning the mouth area, resulted in misalignment of the eyes, eyebrows, and nose, evident in Fig. 8(b), likely caused by differences in position and scale between the faces.

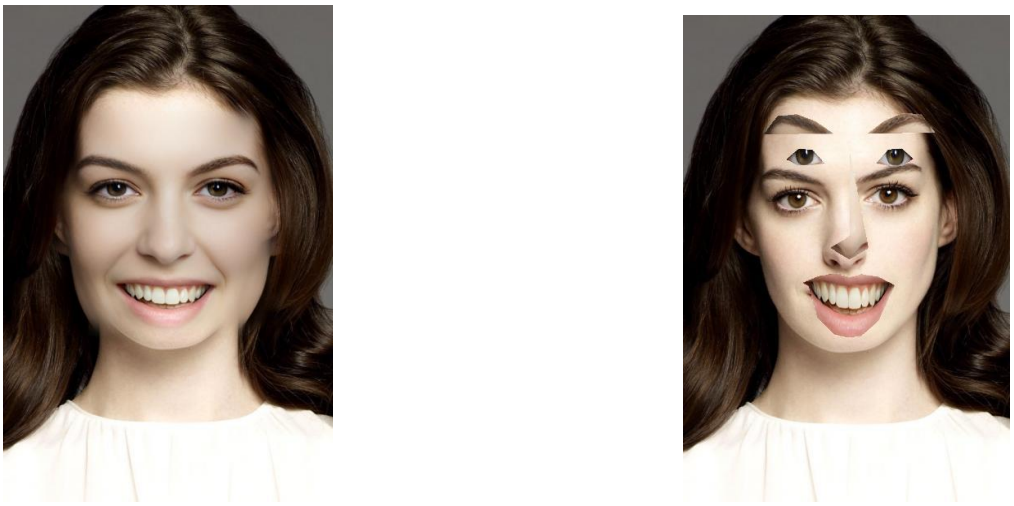


Fig. 8. (a) initial approach of pasting the cropped face to the original image, (b) second approach of paste-back operation.

These failures led to our final approach. Our final "paste-back" operation involves several steps:

1. Identifying facial landmarks in both the processed and original images using the dlib library.
2. Aligning faces based on eye angles to ensure consistent orientation.
3. Extracting the facial region from the cropped image and overlaying it onto the original image.
4. Using soft masks and Gaussian blur to create a seamless transition between the inserted face and the original image's background.
5. Employing OpenCV's 'seamlessClone' function to flawlessly blend the cropped face back into the original image.

This process ensures accurate head movement by restoring the complete context around the face, as shown in Fig. 9.

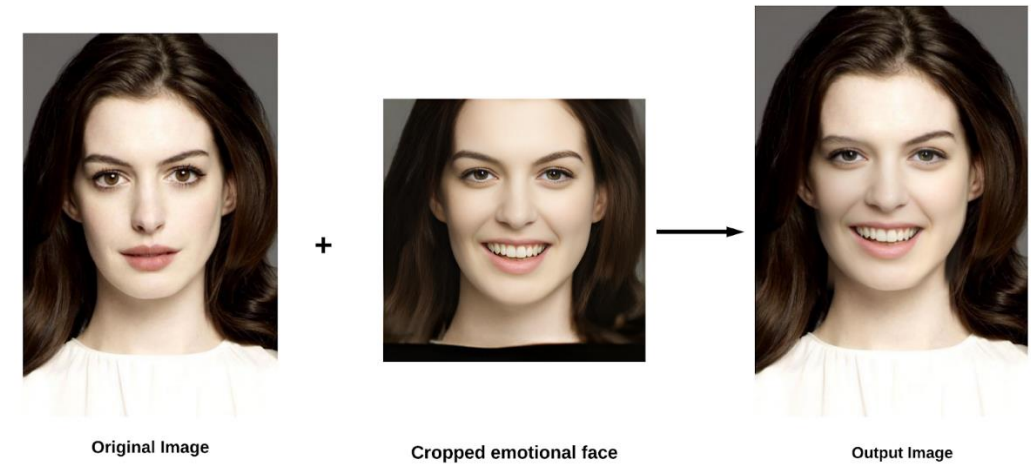


Fig. 9. Final paste-back operation results

2.2.4. Head pose implementation

The head movement plays a significant role in naturalism. However, neither our emotion generator nor our lip synchronisation generator contributes to this element. Therefore, we decided to incorporate a distinct head pose implementation process into our system to address this deficiency.

In talking face generation, the head movement can be generated using two main approaches: one that employs a pose source video and another that learns pose motions directly from the audio input. The study [22] utilizes a deep learning architecture to predict facial landmarks and head poses from speech signals alone. Conversely, studies [23] and [17] depend on pose source videos to adjust head motions, with [17] emphasising the challenges of deducing head pose solely from audio signals and suggesting the use of a pose sequence from a training video clip for better accuracy.

The methods for incorporating pose from a source video have evolved, with early works like FOMM [24] and DVP [25] focusing on either 2D-based methods or extracting 3D Morphable

Models (3DMMs). However, these methods are often subject-dependent and lack generalizability, limiting their applicability in diverse real-world situations.

Addressing these limitations, [26] introduced the Disentanglement of Pose and Expression (DPE) model, a self-supervised framework that separates pose and expression in the latent space without relying on paired data or predefined 3DMMs. This innovative model enhances flexibility and generalises well to a variety of faces, making it a preferable choice for our final approach to generating realistic talking faces. The DPE's Motion Editing module, in particular, stands out for its capacity to produce an edited latent code with multiscale feature maps of the source image, providing a more nuanced and adaptable head movement generation.

In our project, we initially attempted to integrate head pose movements before implementing lip synchronisation. This approach, however, led to issues with the system's ability to accurately recognise and animate the mouth, as the prior head pose adjustments seemed to disturb the mouth's spatial orientation. This is depicted in Fig. 10(a), where the mouth moves without opening. Recognising this problem, we revised our strategy, applying head pose adjustments after the lip synchronisation process. This alteration, shown in Fig. 10(b), preserved the accuracy of the lip movements and produced more satisfactory results, with the mouth movements correctly matching the audio source.

Fig. 10. (a) initial approach of implementing head pose first (the mouth never appears open), (b) final approach of implementing head pose as a post-processing procedure.

2.2.5. Visual quality improvement

The Codebook lookup transformer (Codeformer) [31] presents a unique method for improving face restoration from low-quality inputs using a three-stage process. The first stage involves training an autoencoder with vector quantisation to create a codebook that correlates poor inputs with enhanced outputs. Following this, a Transformer module and a Controllable feature transformation module are used to predict and refine corrupted or incomplete areas, improving results even with varying degrees of input degradation.

In our project, the lip synchronisation stage results in reduced overall image quality. Despite Sadtalker [21] employing multiple loss functions, including Adversarial, Reconstruction, Perceptual, and Expression Loss, the output still contained noticeable noise and flaws. To enhance visual quality, rather than complicating the system with an extra loss function, Codeformer was utilised, a tool designed to identify high-quality facial regions and use this data to improve lower-quality areas of

the face. This technique effectively increased resolution, particularly in blurry areas like the mouth region.

3. Results and discussion

3.1.Emotion

The data in Fig.6 illustrates that while emotions like happiness and disgust are easily discernible in images, it is more complex to distinguish emotions such as anger, sadness, and fear. In the academic community, there is an ongoing discussion about the most representative emotion model, be it categorical or dimensional. Our current model operates on the premise of the categorical emotion model, implying a set number of inherent emotions in humans. This is supported by the CREMA-D dataset [27], which captures 91 actors demonstrating all of Ekman's defined emotion categories, encapsulating six primary emotions. In contrast, the dimensional model posits that emotions are interlinked, with each category being a mix of values from emotional scales. These scales assess the activity and positivity or negativity of an emotion. Adapting to this perspective, our model could emphasise the varying intensity of emotions.

3.2.Lip synchronisation

Our evaluation strategy included both quantitative and qualitative methods. Quantitatively, we achieved superior results in several metrics, scoring an MSE of 53.894, an SSIM of 0.967, and an FID of 18.042, but fell behind in PSNR, ranking third with a score approximately 2.7 points lower than the video-retalking method. MRE quantifies the average error between the generated video and a real source video, PSNR assesses the quality of reconstructed images or videos, and SSIM measures the visual impact of errors in the structural information. In addition, FID measures the distance between the data distribution of the generated output and the real source.

This assessment involved comparisons with cutting-edge methods like PC-AVS [23], Video-retalking [20], and SPT Wav2Lip [31] as detailed in Section 3.2 and illustrated in Table 1.

Table 1: lip synchronisation quantitative evaluation.

	PSNR ↑	MSE ↓	SSIM ↑	FID ↓
SPT Wav2Lip	33.791	82.680	0.945	19.609
PC-AVS	14.985	6197.761	0.453	222.123
Video-retalking	**34.958**	62.498	0.955	74.156
Sadtalker+ Codeformer	32.245	**53.894**	**0.967**	**18.042**

For the qualitative evaluation, we conducted a survey with 10 participants aged 18 to 50 to gain diverse insights on lip synchronization quality. Using a 0-10 scale, our method, 'EmoFaceGen', scored 6.7, outperforming the SPT Wav2Lip's5.7. The 'Ground truth', which is the real source video, scored 9.1. These results, displayed in Fig. 4, show a clear difference between real and generated videos but also highlight the advancements our approach has made, especially regarding

the precision of lip shapes, as recognised by the participants. The survey results are shown in

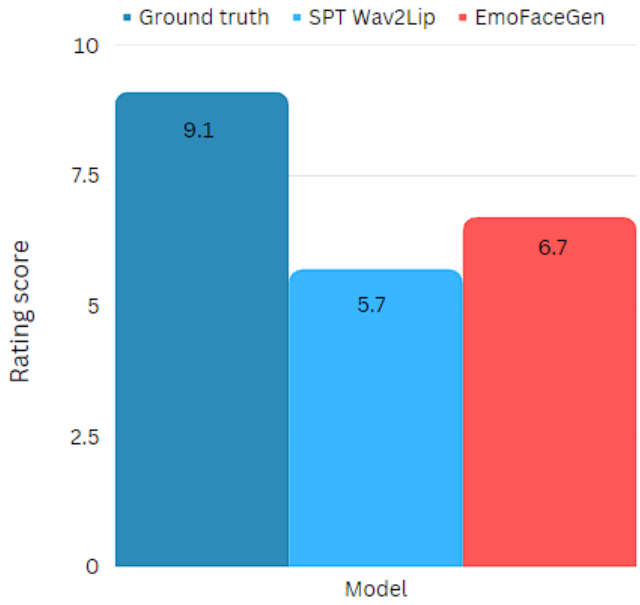


Fig. 11. Lip synchronization quality survey results.

3.3. Overall system from steps 2 to 5

A thorough qualitative analysis was undertaken, from steps 2 through 5, to evaluate the integration of lip synchronization, head pose dynamics, and visual enhancements. Similar to the lip synchronization evaluation, a survey asked participants to rate the realism of videos, both real videos and those rendered by our system, on a scale of 0 to 10. In this assessment, our methodology achieved a 6.0, whereas the ground truth videos scored 9.1, as depicted in Fig. 12. This gap might be due to slight unnatural elements introduced while adjusting the head pose and enhancing image quality. However, the score of 6.0 received by our produced videos indicates a leaning towards a more positive reception.

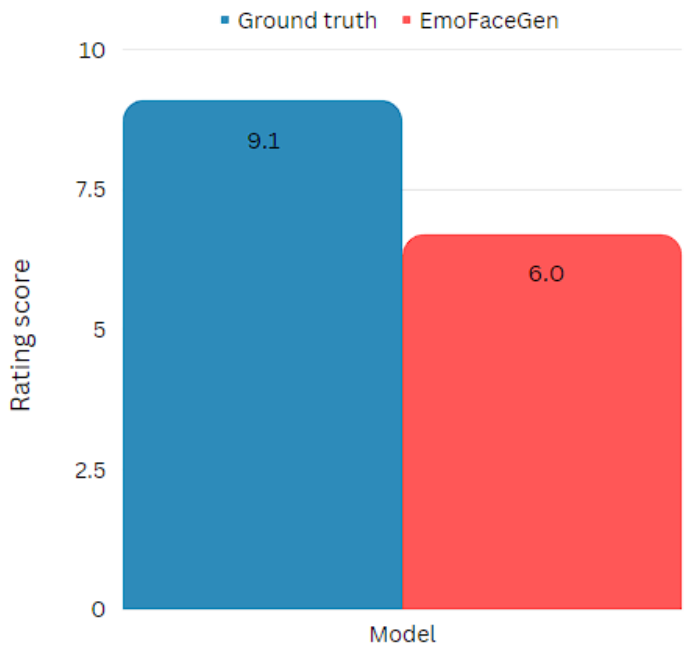


Fig. 12. Survey results for a combination of lip synchronisation and post-processing procedures quality.

4. Instructions to set up environment

The development was conducted using the university's resources, including a specific university lab computer and a Docker environment equipped with a Quadro P6000 GPU and 24576MiB of RAM. We also employed Google Colab Pro, which comes with a T4 GPU and 32GB of RAM. For testing purposes, we prepared Google Colab notebooks to enable easy execution.

4.1.Overall System

The entire system operates through the use of two independent Google Colab notebook files.

The first notebook focuses on emotion generation and paste-back operation. Furthermore, it converts image to video for lip synchronisation generator input.

The second notebook includes lip synchronisation, head pose implementation and visual quality improvement. It takes a video generated from the first file and an audio file path generated from text as inputs.

4.2.Individual components

4.2.1. Emotion

The emotion generation process is carried out using the “<Emotion_generator>.ipynb” Colab notebook located in the "2023_emotion" branch. This script offers two main functionalities: producing emotional images and improving their visual quality. When the script is executed, two folders are generated: "emo_result," which contains emotional images without visual quality enhancement, and "result," showcasing the enhanced-quality emotional images.

4.2.2. Pre-processing

After obtaining an emotional image from the emotion generator, proceed by running it through “<preprocessing_pasteback_video>.ipynb”. This colab notebook is under “2023_lipsync_headpose” branch. This script includes two primary operations. The initial step involves a paste-back operation, as detailed in Section X. Following this, the image is transformed into a video sequence, a necessary format since the lip synchronisation generator requires video input.

4.2.3. Lip synchronisation and head pose implementation

For the execution of lip synchronisation and head pose implementation, there are separate folders. These folders are under “2023_lipsync_headpose” branch. Each folder contains a “requirements.txt” file, providing a list of necessary dependencies. Additionally, there are instructions within the “readme” files to guide users through the dependency installation process. Moreover, the “readme” contains detailed guidelines for the testing procedures. For the integration of lip synchronisation and head movement, a Google Colab notebook “<lipsync_and_headpose>.ipynb” under “2023_lipsync_headpose” branch is provided where they

can be handled simultaneously. Users can also execute lip synchronization and head pose estimation separately, facilitating the ease of reproducing individual modules.

For lip synchronisation process, the user can run PSNR, MSE, SSIM, and FID metrics by running "<lipSync_evaluation>.ipynb in the lip synchronisation folder or under "2023_lipsync_headpose" branch.

5. Seminar & Exhibitions

We participated in a seminar on the 22nd of July, where we had the opportunity to present our project using a series of slides, we had prepared in advance. During this seminar, we received constructive feedback regarding the readability of our presentation; specifically, it was suggested that we increase the font size on our slides to enhance visibility. We have taken this valuable advice into consideration and plan to implement these changes to ensure our presentation is clear and accessible for our final exhibition day. The assessment for our exhibition is scheduled for the 19th at 9:30 AM, and we are committed to incorporating the feedback to improve our presentation effectively. On the day of the exhibition, we will discuss our findings from the literature review and describe the evolution of our implementation, detailing the various stages and outcomes of our system. During the seminar, time constraints prevented us from incorporating the head pose feature; however, by the time of the exhibition, we had successfully completed the entire pipeline for demonstration.

6. Conclusion

This project focused on creating a realistic talking face from a single image and a text-converted audio source. Initially, we aimed to develop a unified model for emotion generation and lip synchronisation but faced issues with unrealistic visuals and prolonged training times. We then shifted to a method that first rendered emotions in the image, followed by lip synchronization to

match the audio, significantly improving realism. However, the lack of head movements reduced the authenticity of the talking face, leading us to integrate head movements post-lip synchronization. An obstacle was the emotion generator's cropping of the face, restricting the head pose module's effectiveness. We solved this by pasting the cropped face back onto the original image before head movement implementation. For enhanced visual quality, we employed a transformer model to improve the image's lower-quality areas using data from higher-quality sections, achieving a more realistic output.

Our method outperformed other state-of-the-art methods through both quantitative and qualitative assessments, indicating its high overall quality and suitability for applications like human-computer interaction or communicating with images of the left loved ones. Despite these advancements, we plan to refine our approach for non-frontal faces and improve realism further. Additionally, we aspire to incorporate emotion-responsive head movements, adding more depth and naturalness to the generated interactions. Furthermore, the system struggled to precisely create images that show different emotions, often failing to distinguish between certain feelings. This issue is made more complex by the current, active discussions in the emotional studies field, which debate the best way to classify emotions: through set categories or varying dimensions. These discussions emphasise the complexity involved in digitally portraying emotions. Moving forward, our efforts will focus on improving our emotional image generator. We plan to shift to a dimensional model, focusing on the intensity of emotional expressions, to enhance the authenticity and breadth of emotions shown by our digital avatars. This advancement is not only about improving visual quality but also narrowing the emotional divide between synthetic representations and authentic human connections.

Name	Contribution
Gayeon Kim	Lip synchronisation, Text-to-Speech, Paste-back Operation, Head pose implementation, Visual quality enhancement, merging code
Yugyeong Hong	Emotion generator, Visual quality enhancement, merging code

7. List of files submitted

In the Compendium folder, we have included these folders/files:

- Exhibition Day Presentation folder: Includes power point slides for final demo.
- Final Report: Includes our final reports.
- Literature Review and Statement of Research: Includes our literature review reports and mid-year progress report.
- Poster: Includes our exhibition poster in pdf format.
- Weekly Presentations: Includes our weekly presentation slides for our meetings. Also includes the plans for our demo and reports.
- Results: Includes our results. Have sub-folders: Emotion, LipSync, Headpose, Visual_Quality_Improvement. Could not include all of them due to big size of images and videos, therefore, added three images/videos for each of these sub-folders. LipSync folder also has one audio file which is generated by google TTS. More results are stored in our google drive.

Our github: https://github.com/UoA-CARES-Student/TalkingFaceGeneration-with-Emotion.git

“2023_all_together” branch is the one being submitted.

“2023_lipsync_headpose” branch includes lip and head pose implementation described in Section 4.

“2023_emotion” branch includes emotion implementation described in Section 4.

Acknowledgements

I would like to thank my project supervisors, Ho Seok Ahn and Trevor Gee for their guidance and support during this research project.

References

[1] MarketsAndMarkets, “Human Machine Interface Market by Product (Hardware (Basic HMI, Advanced PC Based HMI, Advanced Panel Based HMI) and Software (On Premise HMI and Cloud Based HMI), Configuration (Embedded HMI, Standalone HMI) and Region - Global Forecast to 2027”, Aug. 2022, https://www.marketsandmarkets.com/Market-Reports/human-machine-interface-technology-market-461.html

[2] 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

[3] 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

[4] 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

[5] 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

[6] Yang, Yiju & Williams, Andrew, “Improving Human-Robot Collaboration Efficiency and Robustness through Non-Verbal Emotional Communication”, Mar. 2021, doi: 10.1145/3434074.3447191.

[7] R. Kumar, J. Sotelo, K. Kumar, A. de Brebisson, and Y. Bengio, “ObamaNet: Photo-realistic lip-sync from text,” Dec. 2017, [Online]. Available: http://arxiv.org/abs/1801.01442

[8] L. Li et al., “Write-a-speaker: Text-based Emotional and Rhythmic Talking-head Generation,” Apr. 2021, doi: 10.48550/arxiv.2104.07995.

[9] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, “Synthesising obama: Learning lip sync from audio,” in ACM Transactions on Graphics, 2017, vol. 36, no. 4. doi: 10.1145/3072959.3073640

[10] P. K. R, R. Mukhopadhyay, J. Philip, A. Jha, V. Namboodiri, and C. v. Jawahar, “Towards Automatic Face-to-Face Translation,” Mar. 2020, doi: 10.1145/3343031.3351066

[11] S. E. Eskimez, Y. Zhang and Z. Duan, "Speech Driven Talking Face Generation From a Single Image and an Emotion Condition," in IEEE Transactions on Multimedia, vol. 24, pp. 3480-3490, 2022, doi: 10.1109/TMM.2021.3099900.

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

[13] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8789-8797.

[14] 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

[15] 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

[16] 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

[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/2205.15278

[18] 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

[19] S. Zhou, K. C.K. Chan, C. Li, and C.C. Loy, "Towards Robust Blind Face Restoration with Codebook Lookup Transformer," 2022.

[20] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, C V Jawahar, “A Lip Sync Expert Is All You Need for Speech to Lip Generation In The Wild”, Aug. 2020, doi: https://doi.org/10.48550/arXiv.2008.10010

[21] 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”, 2023

[22] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, Dingzeyu Li, “Make it talk: Speaker-Aware Talking head animation”, Feb. 2021, doi: https://doi.org/10.1145/3414685.3417774

[23] 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, doi: https://doi.org/10.48550/arXiv.2104.11116

[24] Aliaksandr Siarohin, Oliver J. Woodford, Jian Ren, Menglei Chai, Sergey Tulyakov, “Motion Representations for Articulated Animation”, Apr. 2021, doi: https://doi.org/10.48550/arXiv.2104.11280

[25] Xin Wen, Miao Wang, Christian Richardt, Ze-Yin Chen, and Shi-Min Hu, “Photorealistic audio-driven video portraits”, Dec. 2020, doi: https://doi.org/10.1109/TVCG.2020.3023573

[26] Youxin Pang, Yong Zhang, Weize Quan, Yanbo Fan, Xiaodong Cun, Ying Shan, Dong-ming Yan, “DPE: Disentanglement of Pose and Expression for General Video Portrait Editing”, Mar. 2023, doi: https://doi.org/10.48550/arXiv.2301.06281

[27] 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

[28] 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, doi: https://doi.org/10.48550/arXiv.2211.14758

[29] Takahiro Ishiguro, Alan Lin, Jong Yoon Lim, Trevor Gee, Edmond Liu, Bruce A. MacDonald, and Ho Seok Ahn, “Talking Face Generation from Facial Image, Target Emotion and Speech Dialogue”, In Proceedings of the 2022 Australasian Conference on Robotics and Automation (ACRA 2022), Dec. 2022.

[30] Prajwal K R, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, C. V. Jawahar, “Towards Automatic Face-to-Face Translation”, Mar. 2020, doi: https://doi.org/10.1145/3343031.3351066

[31] Shangchen Zhou, Kelvin C.K. Chan, Chongyi Li, Chen Change Loy, “Towards Robust Blind Face Restoration with Codebook Lookup Transformer”, Nov. 2022, doi: https://doi.org/10.48550/arXiv.2206.11253

Appendix A

Students may introduce here a proof of a formula used in the report, a piece of software code, additional graphs, template of a schematic diagram, and any other information that might be helpful in describing the topic under investigation.

This section is not mandatory and it may not be needed in many of the reports, especially literature review.