IMPROVED SPEECH SYNCHRONIZED TALKING FACE GENERATION FROM AN IMAGE AND AN EMOTION CONDITION

R Avaneesh, Sanka Anjani, Shaheen Rafiq, Maneesha A Pillai, Anjalin Anna Cherian

Under the guidance of: Dr. Ameer P. M.

Abstract

By granting computers the ability to communicate emotions through facial expressions, human-computer interactions can be made more organic and natural. Human-Computer Interaction (HCI) is crucial because it plays a central role in bridging the gap between humans and technology, making complex systems and devices accessible and user-friendly. As technology continues to advance rapidly, HCI becomes essential for ensuring that humans can interact efficiently, intuitively, and effectively with computers, software, and various digital interfaces. HCI also facilitates the seamless integration of technology into various aspects of daily life, from smartphones and computers to smart home devices and beyond. This study analyses and refines an emotional talking face model that integrates face images, categorical emotions, and speech signals to produce synchronized facial expressions. Incorporating a squeeze-and-excitation network, introducing non-linearity, and enhancing effectiveness led to a significant improvement in the model's performance. A Blind Image Super-Resolution Generative Adversarial Network enhanced video resolution, while multi-task cascaded convolutional neural networks accurately detected and utilized the eye region. Evaluation of the model employed Fréchet Inception Distance score for objective assessment and Mean Opinion Score for subjective evaluation. Results conclusively demonstrate substantial enhancements in accuracy and performance facilitated by these network additions.

Introduction

Communication is an integral part of our daily lives. While verbal communication plays a vital role in conveying messages, it is often accompanied by nonverbal cues, which are essential in communication as they enhance the clarity and understanding of the message and provide context for the words spoken. The evolution of computer vision and deep learning has made it possible for the development of speech models that can provide both acoustic signals and visual cues by generating realistic and emotionally expressive talking faces. Such models have been of high interest as they are becoming invaluable in various domains including entertainment, virtual assistants, speech therapy and human-to-computer interactions.

Robots have started to work in increasingly people-facing roles like tutors, household supporters, and receptionists, among others [1]. As facial expressions play a vital role in human-human interactions (HHI), giving computers the ability to adapt their "facial expressions" based on the situation can go a long way in making human-computer interactions (HCI) more organic and natural. This can be accomplished by affective computing, i.e., making computers more human-like in their observation, interpretation and generation of affect features [2].

During verbal communication, emotions play a crucial role, directly impacting the transmitted message and sometimes causing significant changes in its meaning [3]. Research indicates that predicting emotions solely from speech audio is challenging for untrained individuals [4], and we heavily rely on visual cues to interpret emotions [5]. Therefore, to enhance the authenticity of visual rendering and improve speech communication, it becomes essential for automatic talking face generation systems to accurately represent emotional expressions visually.

One approach to achieve emotional talking face generation involves first estimating the expressed emotions from the speech utterance and then incorporating them into the generated talking faces. However, this approach faces limitations due to the accuracy of speech emotion recognition and lacks the ability to independently control emotional expression in the visual rendering. In this work, we adopt a different approach: we do not consider emotions expressed in the speech audio but instead condition the talking face generation on an independent emotion variable. This allows for direct and flexible control over visual emotion expression, enabling more personalized applications in entertainment, education, and interactive assistive devices. Moreover, it offers a powerful tool for behavioral psychologists to conduct emotion-related experiments that were previously unfeasible. For instance, researchers can now investigate how humans respond to and interact with conversational partners' emotional expressions by manipulating emotions independently in both audio and visual modalities.

The ability to generate realistic facial animations and lip-syncing from a single still image and an audio input has many valuable applications in the fields of entertainment, education, and human-computer interaction. It also has the potential to advance the field of psychology by uncovering new findings related to human behavior and emotions, as it enables more precise and controlled experiments to be conducted that can provide insights into human emotional responses and social interactions. It can also be extended to other applications such as video conferencing, where users can interact with realistic avatars that mimic their facial expressions and speech.

In this paper, we introduce an enhanced version of a neural network system designed to generate emotional talking faces from speech conditioned on categorical emotions. The proposed network takes inputs such as a speech utterance, a reference face image, and a categorical emotion condition, and then produces a talking face that synchronizes with the input speech while expressing emotional cues. Our primary contributions are as follows:

- We suggest incorporating a Squeeze and Excitation-based attention layer to enhance the image encoder.
- We propose the utilization of an eye region mask-based loss to improve the model's performance in the critical eye region.
- We introduce a BSRGAN-based super-resolution module to enhance the overall visual quality of the generated talking faces.

The paper is structured as follows: In Section II, we review related work on talking faces. Section III outlines our proposed method and objective functions. Experimentation details, along with objective and subjective evaluations, are provided in Section IV. Lastly, we summarize our findings and conclude the paper in Section V.

Related Works

A comprehensive review of the available literature in the relevant research fields has been conducted, and the significant related studies are outlined below.

Facial emotion recognition (FER) is challenging due to pose, lighting, and occlusion changes. There are several methods for performing facial emotion recognition, including filtering techniques such as Gabor wavelets, histogram of oriented gradients, and local binary patterns, as well as feature encoding using code blocks and spatial pooling [6]. Due to its vital role, research on face expression has a long history, especially in coding, recognition and generation. The Facial Action Coding System (FACS) classifies emotions into six categories: joy, sadness, disgust, fear, surprise and anger [7].

Upon reading more about this, we came across deep learning for image generation. Additionally, we encountered the term Generative Adversarial Networks (GAN) [8] and gained a foundational understanding of it [9]. An increasing number of efficient GAN architectures have been developed and suggested to effectively learn the diverse variations of human faces, including cross pose, age, expression, and style. The development of GANs has led to the development of many new architectures which can translate from one image domain to another. The two main GAN architectures in this domain are conditional GAN (cGAN) and cycle GAN. M. Mirza and S. Osindero introduced the concept of conditioning GANs on additional information to control image generation [10]. P. Isola et al. then proposed a cGAN architecture which can perform various image-to-image translation tasks such as style transfer, colourisation and segmentation [11]. Another popular Image to Image (I2I) model is cycle GAN first introduced by P. Isola et al in [12]. It is a method for unsupervised image-to-image translation that learns to map an image from one domain to another without requiring paired examples. These GAN models can be used for facial expression manipulation as we can map from one emotion to another [13]. GANs have come a long way from generating grayscale images to realistic images with suitable levels of style control. They leverage the concept of feature detection by allowing the generator to learn and reproduce essential characteristics of the training data, resulting in the generation of realistic and high-quality synthetic data.

Face detection, identification of key points, and feature extraction are essential for any face-related task. The authors of [14] proposed an innovative method for face feature extraction capable of dealing with variations in pose, using a CNN on a pre-detected face. Face detection can be accurately done using the model proposed in [15]. The MTCNN model can detect and align faces in images. It has three subnetworks: Proposal, Refinement, and Output networks. The subnetworks use convolutional and max-pooling layers. The Proposal network (P-Net) further uses a dense layer to obtain probability maps. The Refinement Network (R-Net) filters false positives from P-Net and gives bounding boxes. The Output network (O-Net) further refines the bounding boxes, resulting in a robust model which can detect and align faces accurately. Detection of key features of the face can help draw the attention of the generator to specific features [16]. This can help make the model more robust and improve the output in that specific region. When using an alpha channel or "mask," individuals can selectively edit or apply effects to specific portions of an image. This capability can be leveraged when training a GAN model to create or modify the unmasked region solely. The process involves constructing a mask covering the entire image except for the designated area of interest, which can then be used to specify the region the model should focus on. A loss function can then be designed to penalize the model for producing images that deviate from the desired region's desired attributes [17].

The output of Convolutional layers can be enhanced using squeeze and excitation layers. The main idea is to learn to "squeeze" the channels of feature maps to capture the most important features and then "excite" them by learning to weigh the channels dynamically. For this, we utilize a mechanism called channel attention, which assigns importance scores to different channels and uses them to scale the feature maps [18].

Another way to improve the output generated by GAN models is to enhance them by means of super resolution, which is to recover a high-resolution image from a single low-resolution image. Enhanced Super-Resolution Generative Adversarial Networks (ESRGAN) is a cutting-edge deep learning technique for single image super-resolution. Introduced in [19], ESRGAN has become a prominent method in the field by employing a generative adversarial network framework to generate highly detailed and visually pleasing high-resolution images from low-resolution inputs. By incorporating residual blocks and a perceptual loss function, ESRGAN excels in capturing fine details and ensuring perceptually similar outputs, surpassing traditional interpolation-based methods and earlier deep learning approaches. ESRGAN continues to be an active area of research, with ongoing efforts to enhance its capabilities and further improve the quality of super-resolved images. The authors of [20] provide a complex but practical degradation model that consists of a randomly shuffled blur, downsampling and noise degradations enhancing ESRGAN's outputs. The new degradation model proposed can cover a wide range of degradations found in real-world scenarios.

Furthermore, expression manipulation using deep learning techniques has emerged as a captivating research area, offering the ability to edit and transform facial expressions in images and videos. In the realm of expression editing, existing methods can be classified into two main categories. The first category focuses on manipulating images by reusing parts from existing ones. This approach involves techniques that extract facial components or features from different images and combine them to create a new image with the desired expression [21], [22]. Early techniques generated new expressions using fully textured 3D facial models, face image warping through feature correspondence and optical flow, or compositing face patches from existing expression datasets. While these methods often produce high-resolution and realistic images, their complex processes can be computationally expensive. The second category of methods resorts to synthesis techniques to generate a facial image with the target expression. These techniques leverage generative models to synthesize new images that exhibit the desired expression while preserving other facial attributes [11], [12]. However, the images generated by these methods may lack fine details, appearing blurry or of low resolution. Expressional attributes are typically encoded in a latent feature space, aligning specific directions with semantic properties. While this offers better flexibility in semantical-level image generation, it becomes challenging to precisely control fine-grained aspects of the synthesized images, such as adjusting the degree of a smile or narrowing the eyes.

Hui Ding, Kumar Sricharan, and Rama Chellappa were the first to propose Expression GAN, a novel model that utilises a combination of cGAN and Adversarial Autoencoders (AAE) to enable the editing of facial expressions [23]. The development of computer-generated visuals, specifically facial animations and virtual avatars has undergone a significant evolution over the past few decades, especially with the convergence of computer vision, machine learning and deep learning. Several studies have been conducted in recent years on generating talking faces from audio inputs. Some researchers have used deep learning models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to generate facial animations from speech [24].

Other studies have focused on the idea of rendering the emotions in generated talking faces by estimating them directly from the speech input [25] which has major drawbacks such as limitation in the accuracy of recognition of the emotion and restriction of direct influence or control of the emotion to be rendered in the output. These problems can be solved by the approach we follow, which involves ignoring the emotions expressed in the speech audio and instead conditioning the talking face generation on an independent emotion variable. This has the potential to revolutionize communication across various industries however, achieving realistic and synchronized facial animations with speech has been a persistent challenge

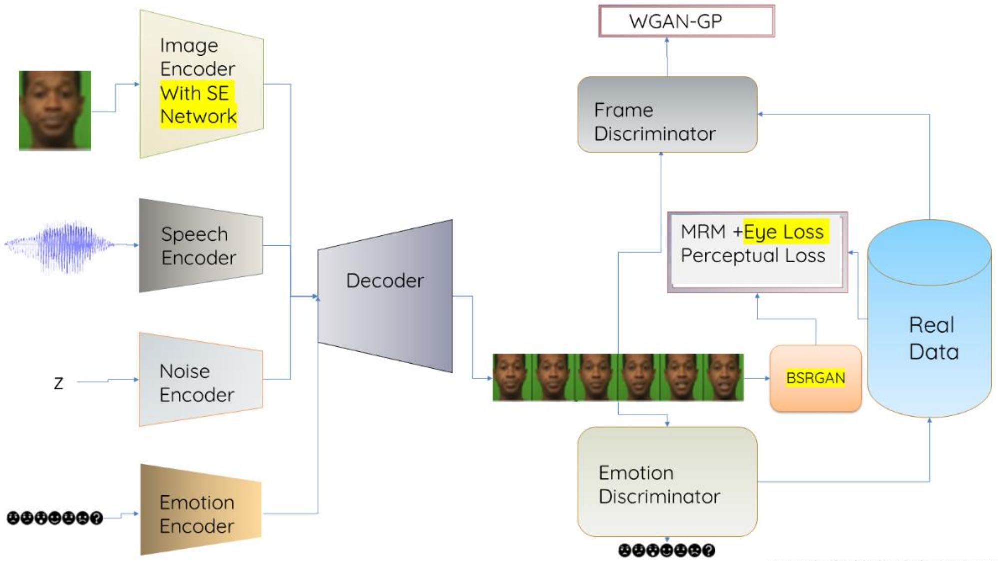
A novel approach for the end-to-end generation of talking faces was proposed by Eskimez et al. in [17]. This model was further refined to generate talking faces from a single image, speech signal and emotion condition by the authors in [26]. The fundamental objective of this paper is to explore how artificial intelligence and deep learning can potentially be utilized to make more realistic and alluring virtual characters, ultimately enhancing user experience and human-computer interaction.

The proposed network aims to leverage recent advancements in deep learning, specifically generative adversarial networks and emotion recognition, to implement and enhance the model for generating realistic talking faces from a single image and an audio input since even though there has been significant progress in speech synthesis and facial animation, generating such high-quality synchronized talking faces remains a daunting task. A novel approach for generating talking faces using a single reference face image and a categorical emotion condition along with the required speech utterance as inputs, leveraging a deep learning framework that utilizes a Generative Adversarial Network (GAN) has been analyzed and implemented in the proposed work. Further we have enhanced the model's performance by integrating additional architectures and networks such as Multi-Task Cascaded Convolutional Neural Networks (MTCNN), Blind Image Super-Resolution GAN (BSRGAN), and Squeeze and Excitation (SE) networks.

Method

The neural network system that we have implemented, which was proposed in the [26] to generate a video with a talking face from a said emotion and an image is as follows:

The system employs GAN architecture which consists of a generator and a discriminator. The inputs are speech waveform and reference image under preconditioned settings. Along with this, we have objective functions to evaluate the losses incurred by the network. The generator and the discriminator can be further broken down into sub-networks.



Note: The highlighted elements represent all the additional networks added to improve the model.

Figure 1 Block Diagram of the model

Generator

The generator network comprises the following sub-networks: speech, image, noise, emotion encoders, and a video decoder.

Speech encoder

The input speech waveform is taken by the speech encoder, and a speech embedding is produced as output. The network is comprised of five 1-D convolutional layers operating in the time domain. The kernel sizes, number of filters, and strides for these layers are (63, 64, 4), (31, 128, 4), (17, 256, 2), (9, 512, 2), and (1, 16, 1), respectively, with each layer being followed by a Leaky ReLU activation having a 0.2 slope. 8 kHz speech signals are accepted by the network, generating 125 feature vectors for each second of speech. A context layer is added after these five convolutional layers to concatenate past and future speech features. The number of time steps is reduced from 125 to 25 by only passing every fifth frame to the next layer, resulting in a generated video with 25 frames per second (FPS). The output of the context layer is then passed through a fully connected layer and two LSTM layers, which produce the speech embedding sequence.

The speech encoder takes the input speech waveform and produces a speech embedding as output. The network comprises five 1-D convolutional layers that operate in the time domain. The kernel sizes, number of filters, and strides for these layers are (63, 64, 4), (31, 128, 4), (17, 256, 2), (9, 512, 2), and (1, 16, 1), respectively, and each layer is followed by a Leaky ReLU activation with a 0.2 slope. Since the network accepts 8 kHz speech signals, it generates 125 feature vectors for every speech second. We add a context layer after these five convolutional layers to concatenate past and future speech features. The context layer reduces the number of time steps from 125 to 25 by only passing every fifth frame to the next layer, resulting in a generated video with 25 frames per second (FPS). The output of the context layer is then passed through a fully connected layer and two LSTM layers, which produce the speech embedding sequence.

Image Encoder

The image encoder processes the input condition image to generate an image embedding.

The architecture is comprised of six layers of 2-D convolutional layers with specific values for the number of filters, kernel sizes, and downsampling factors, namely (64, 3, 2), (128, 3, 2), (256, 3, 2), (512, 3, 2), (512, 3, 2), (512, 4, 1). Each convolutional layer is followed by a LeakyReLU activation with a 0.2 slope introducing non-linearity to the network. Following this, there is a squeeze and excitation block.

The SE network can be broken into two major parts: squeeze and excitation [18]. In the Squeeze process, the spatial dimensions of the feature map are compressed by the network. For this purpose, an adaptive average pooling layer has been used to perform a global average pooling operation, providing a channel-wise descriptor that captures the relevance of each channel. A sequential container of four layers has been defined for the excitation process, where the channel-wise descriptor is taken as input, and a set of scaling factors for each channel is produced. This includes two fully connected layers with a ReLU activation function layer incorporated between them, and the final layer is the sigmoid function layer to obtain the scaling factors. The scaling factors are then expanded to have the same spatial dimensions as the input feature map. An element-wise multiplication of the modified scaling factor is performed with the input feature map to recalibrate the channel-wise features, thus forming the output of the SE block.

After the SE block, dropout is applied to regularize the network by randomly setting a fraction of the input channels to zero during training.

Emotion encoder

The emotions are labelled using a one-hot vector and fed into the emotion encoder. The emotion encoder applies a two-layer fully connected (FC) neural network to project the one-hot vector onto an emotion embedding. This embedding is duplicated for each step. Once more, we employ a LeakyReLU activation with a slope of 0.2 after every FC layer.

Noise Encoder

The noise encoder generates a noise vector from the Gaussian distribution for each video frame. A single-layer LSTM (Long Short Term Memory) processes this sequence of noise vectors and outputs the noise embedding.

Decoder

Here, the decoder takes in the speech, image, noise, and emotion embeddings altogether as input. At every time step, the decoder employs convolutional layers to project the embeddings onto 4 x 4 images via two FC layers and reshape operations. These 4 x 4 images are merged channel-wise with the skip connections arising from the image encoder in the U-Net style for the upcoming layers, except for the last layer. The number of filters in each convolutional layer is identical to the corresponding layer in the image encoder. A LeakyReLU activation with a slope of 0.2 is applied after each convolutional layer, except for the final layer. Instead, a hyperbolic tangent activation is utilised for the last layer since the images are normalised to have values ranging from -1 to 1.

Frame Discriminator

The frame discriminator seeks to increase the visual quality of the produced video while still maintaining the target identification throughout the film. Initially, we duplicate the target image to match the number of frames in the input video and combine them. Afterwards, each frame goes through five layers of 2-D convolutional layers, each with varying numbers of filters, kernel sizes, and strides, as follows: (64, 3, 2), (128, 3, 2), (256, 3, 2), (512, 3, 2), (512, 3, 2), respectively. The outcome is then smoothed and given to a two-layer FC network, which classifies the frame as genuine or counterfeit. Each layer in the FC network is accompanied by a LeakyReLU activation with a 0.2 slope, except for the last layer.

Emotion Discriminator

The emotion discriminator is a video-based emotion classifier with the addition of a false video class. Its goal is to enhance the emotional expression produced by our network. The initial section of the network utilises the same framework as the frame discriminator, comprising five layers of a 2-D convolutional layer trailed by two fully connected layers. During the discriminator's training step, we compute the sparsely categorised cross-entropy loss using the genuine video's emotion label and the fabricated video's fake label. While updating the generator, we process every video frame and input the resulting sequence into an LSTM layer. The ultimate time step of the LSTM layer's output is entered into an FC layer, which generates probabilities for the seven classes: six emotions (anger, disgust, fear, happiness, neutral, and sadness) along with the fake category. We compute the sparse categorical cross-entropy loss employing the emotion label utilised to create the video.

Image Super-Resolution Module

BSRGAN (Blind Super-Resolution with Generative Adversarial Networks) is a type of deep learning algorithm that uses a generative adversarial network (GAN) to generate a high-resolution image from a low-resolution image. The algorithm can do this without having any prior knowledge about the low-resolution image, such as its scale or degradation type. This is why it has 'blind' in its title [20].

The BSRGAN algorithm consists of a generator network and a discriminator network. The generator network takes a low-resolution image as input and tries to generate a high-resolution image that is like the ground truth high-resolution image. The discriminator network will try and distinguish between the generated high-resolution images and the ground truth high-resolution images.

Table 1: Model Summary

Module	Input	Layers	Output
Encoder			
Speech Encoder	Input waveform	1-D convolutional layers, context layer, LSTM layers	Speech Embedding sequence
Image Encoder	Input Image	2-D Convolutional (x6) with LeakyReLU and Squeeze and excitation netwrok	Image Embedding
Noise Encoder	Gaussian noise vector for each video frame	Single layer LST	Noise Embedding
Emotion Encoder	One hot emotion vector	Two-layer FC Neural Network with LeakyReLU	Emotion Embedding
Decoder			
Video Decoder	Speech, image, noise, and emotion embedding	Convolutional Layers with skip connections	Generated Video
Image Super Resolution			
BSRGAN	Generated Video	Pretrained BSRGAN model	Higher Resolution Generated Video

Discriminator			
Frame Discriminator	Video Frames (Real or Generated)	5 Convolutional Layers and 2 Fully Connected Layers	Real or Fake classification for each frame
Emotion Discriminator	Generated Video	5 Convolutional Layers and 2 Fully Connected Layers	Classification of emotion expression

Objective Function

Mouth Region Mask Loss

MRM (Mouth Region Mask) Loss: This loss function focuses on the mouth region of the generated face. It encourages the model to generate accurate and realistic mouth movements that synchronize with the input speech. The MRM loss penalizes any discrepancies between the generated mouth region and the ground truth mouth region.

Perceptual Loss

Perceptual loss is a commonly used loss function in image generation tasks. It measures the difference between the features extracted from the generated image and the features extracted from a reference image. By minimizing this loss, the model learns to generate visually similar faces to the reference image, ensuring that the generated faces resemble the original face as closely as possible.

Eye Region Mask Loss

Similar to the MRM loss, the Eye Region Mask loss focuses on the eye region of the generated face. It encourages the model to generate accurate and realistic eye movements that convey emotions.

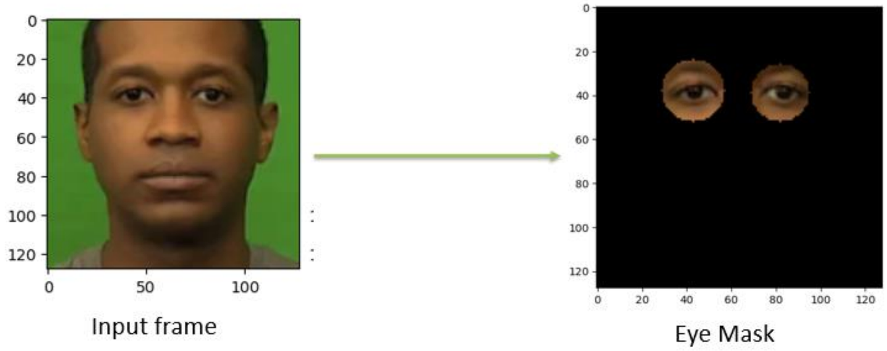


Figure 2:Implementation of Eye mask on a single frame

By penalizing discrepancies between the generated eye region and the ground truth eye region, the model learns to generate expressive and emotionally consistent eye movements.

Frame Discriminator Loss

The frame discriminator loss is associated with a discriminator network that distinguishes between real and generated frames in a sequence. By training the model to minimize this loss, the generator becomes more adept at producing frames that are indistinguishable from real frames, thereby enhancing the overall realism of the generated talking face animations.

Emotion Discriminator Loss

In addition to the frame discriminator, the emotion discriminator loss is related to a discriminator network that distinguishes between different emotion conditions. This loss function encourages the generator to generate faces that accurately convey the specified emotion condition. By minimizing this loss, the model becomes proficient at generating faces with the desired emotional expressions.

The Complete objective function is as follows:

$$J_{GEN} = \alpha L_1^{MRM} + \beta\, L_2^{Percptual} + \gamma\, J_{FD} + \delta J_{ED} \; + \epsilon \mathrm{L}_1^{\mathrm{EL}}$$

In this equation, J_{GEN} epresents the generator loss, $L_1^{MRM}$is the MRM loss, $L_2^{Percptual}$ is the perceptual loss, J_{FD} is the frame GAN loss, J_{ED} is the emotion GAN loss, L_1^{EL} represent the calculated eye loss and α, β, γ, δ and ϵ are the respective weights of each component. We have taken $\alpha = 100, \beta = 1, \gamma = 0.01, \delta = 0.001$ and $\epsilon = 1$.

Experiments

Dataset

Crowd-sourced Emotional Multimodal Actors Dataset or CREMA-D [27], is the dataset that has been used. It is a publicly available database containing 7,442 original clips from 91 actors, specifically 48 male and 43 female actors between the ages of 20 and 74 coming from a variety of races and ethnicities. The actors were made to speak short phrases from a set of 12 sentences while displaying a wide range of emotions, specifically anger, disgust, fear, happiness, neutral and sadness. These videos are sampled at 30 Frames per second (FPS), while the speech is sampled at 44.1 kHz. The videos have an image resolution of 480x360. Crema-D has been widely used in the research community to develop and evaluate models for various tasks such as facial expression recognition, emotion recognition, and audio-visual speech processing.

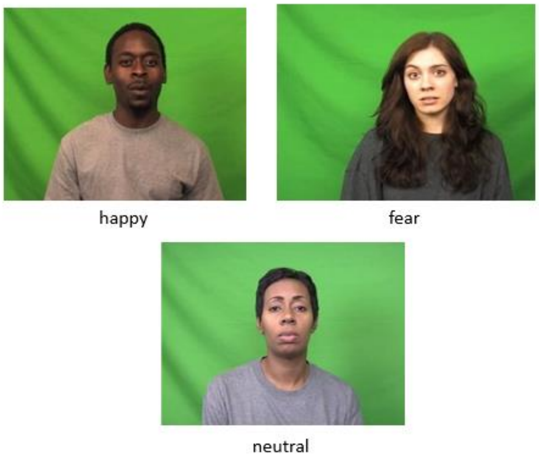


Figure 3: CREMA-D Selected Samples

The pre-processing of the dataset used in the model includes downsampling the video to 25 FPS and audio to 8 KHz to reduce computational complexity. Further, to facilitate a better training process, the actor's faces was aligned across videos by extracting facial landmarks from a chosen template image and extracting landmarks of each video's first frame.

Implementation Details

The Dataset we used for this model is CREMA-D. We only used a portion of this dataset due to memory constraints for our training. So, we used a total of 100 videos for training. To train the model, we have utilized an incremental training approach as it gives better results [28]. We have trained it in two stages; the first stage was trained to calculate MRM loss and perceptual loss. The second stage of training was done with the complete objective function. We have trained both the discriminator and the generator for 500 epochs each and 200 epochs together. We utilized a batch size of 2 for the discriminator, 2 for the generator, and 2 when training both together. We have used Adam optimiser for all the networks, and the learning rate for the generator was 1e-4 during the initialization and 1e-5 during the GAN training. Both discriminator's learning rates were 1e-4. The implementation was conducted on a P100 GPU.

We have also implemented the above model using the DGX machine. Here, the generator and the discriminator were trained for 1600 epochs each and 2000 epochs together. We have used Adam optimiser for all the networks, and the learning rate for the generator was 1e-4 during the initialization and 1e-5 during the GAN training. Both discriminator's learning rates were 1e-4. It was conducted on A100 GPU. We utilized a batch size of 2 for the discriminator, 2 for the generator, and 2 when training both together for this as well.

Results

Below are the improvements obtained by incorporating the various modules and enhancements.

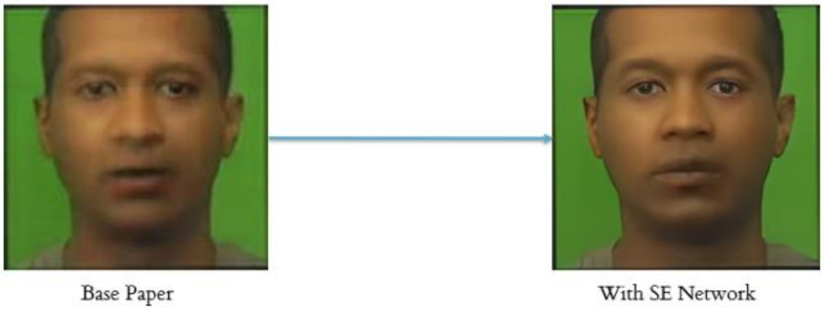


Figure 4: Improvement obtained by using SE Network

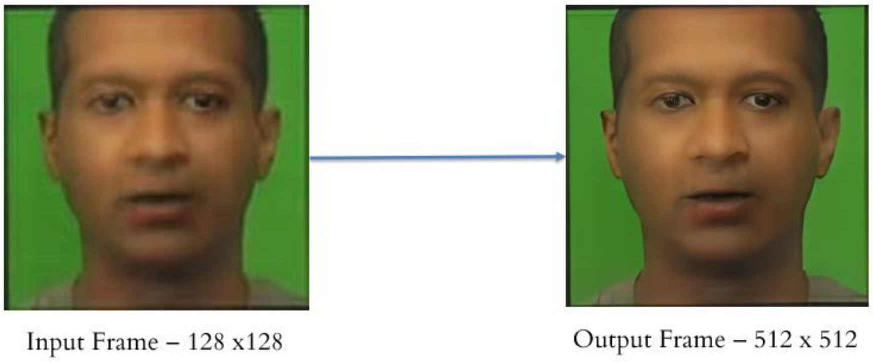


Figure 5 : Improvement obtained by using Super Resolution Module

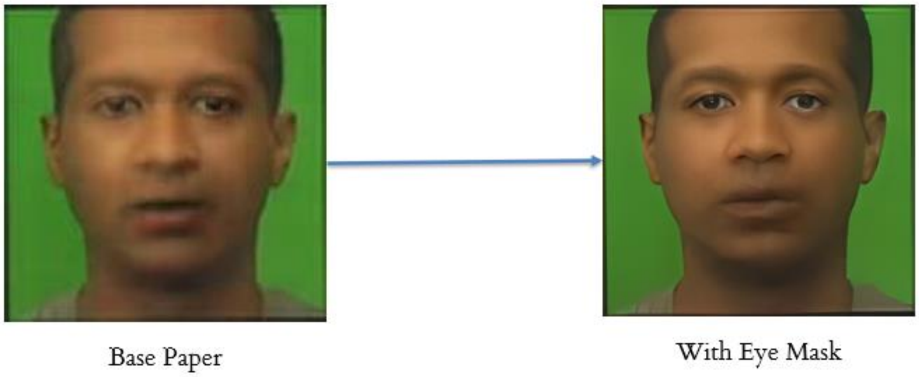


Figure 6:Improvement obtained by using Eye Mask

Below are snapshots of a particular frame from the outputs obtained from different models:

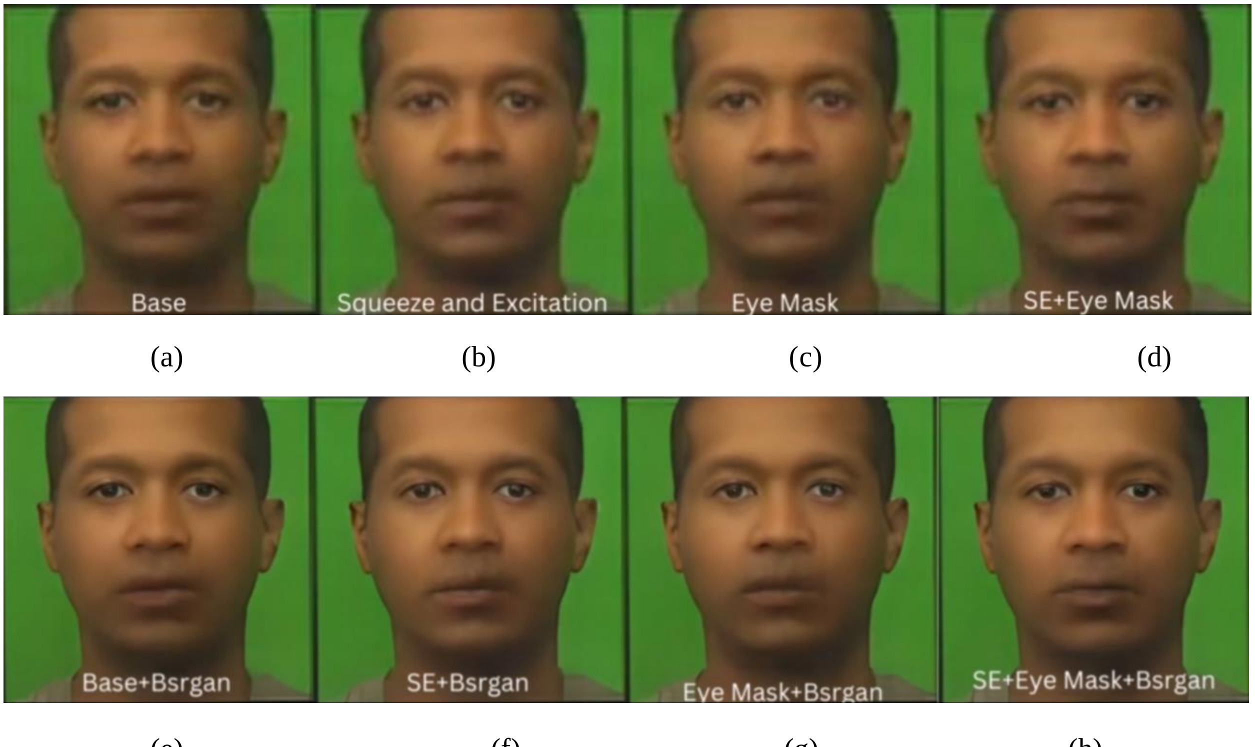


Fig. 7 Output of Different models. (a)Base Model, (c) Squeeze and Excitation Model (c)Eye Mask Model (d) SE + Eye Mask Model (e) Base + Bsrgan Model (f) SE + Bsrgan Model (g) Eye Mask + Bsrgan Model, (h) SE + Eye mask Model+ Bsrgan Model

From the results of the different models as shown in Fig. 7, it can be observed that incorporating the SE network and eye mask improves the quality of the output in the mouth and eye regions. Furthermore, adding BSRGAN significantly improves the resolution. Finally, the model with all three components (SE, eye mask, and BSRGAN) produces the best results.

The model results for different emotions at the 1600, 1800, and 2000 epochs are shown in Figure 8 . It was observed that some emotions, such as anger, fear, and neutrality, provided better results than others (Figure 10).

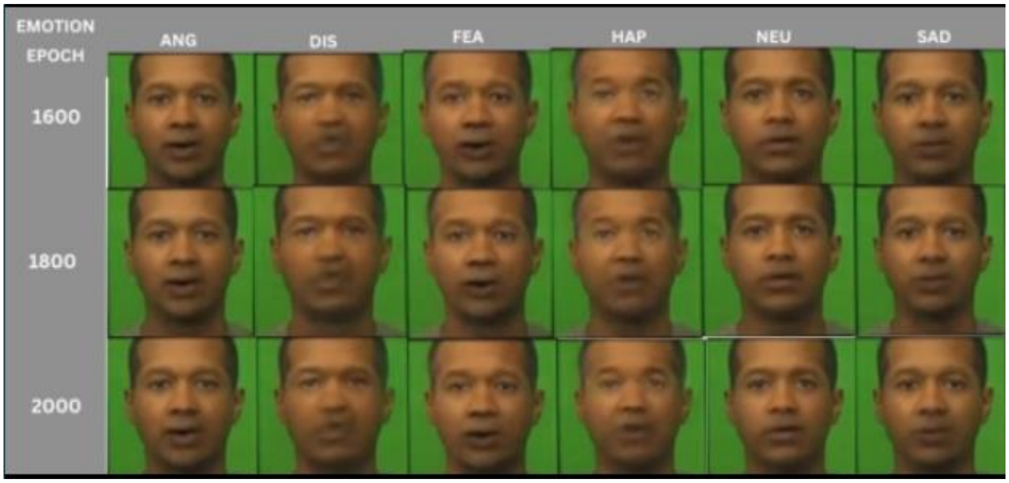
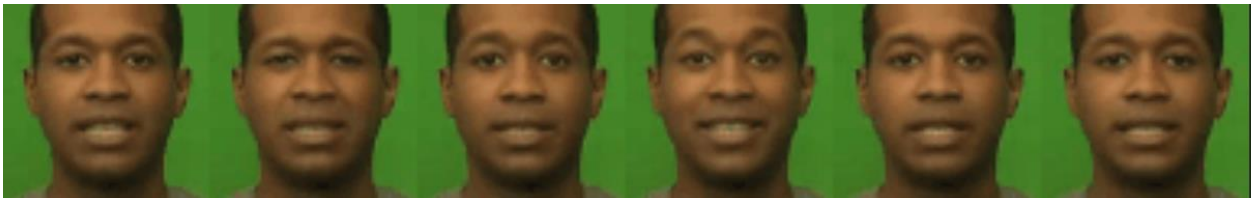
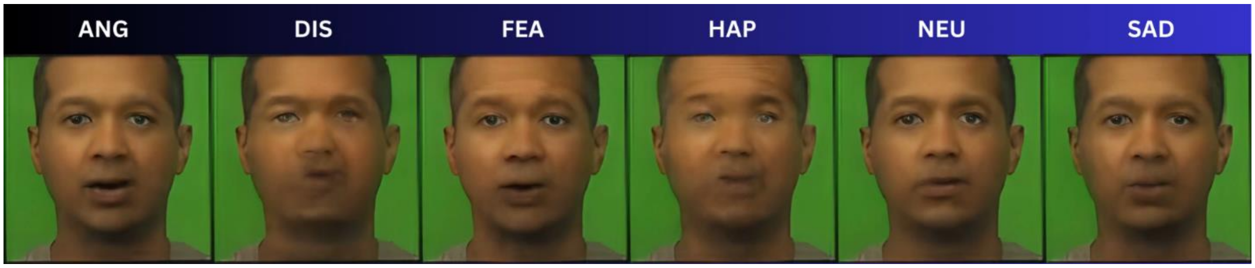


Figure 8 Output with different epochs



(a) (b) (c) (d) (e) (f)

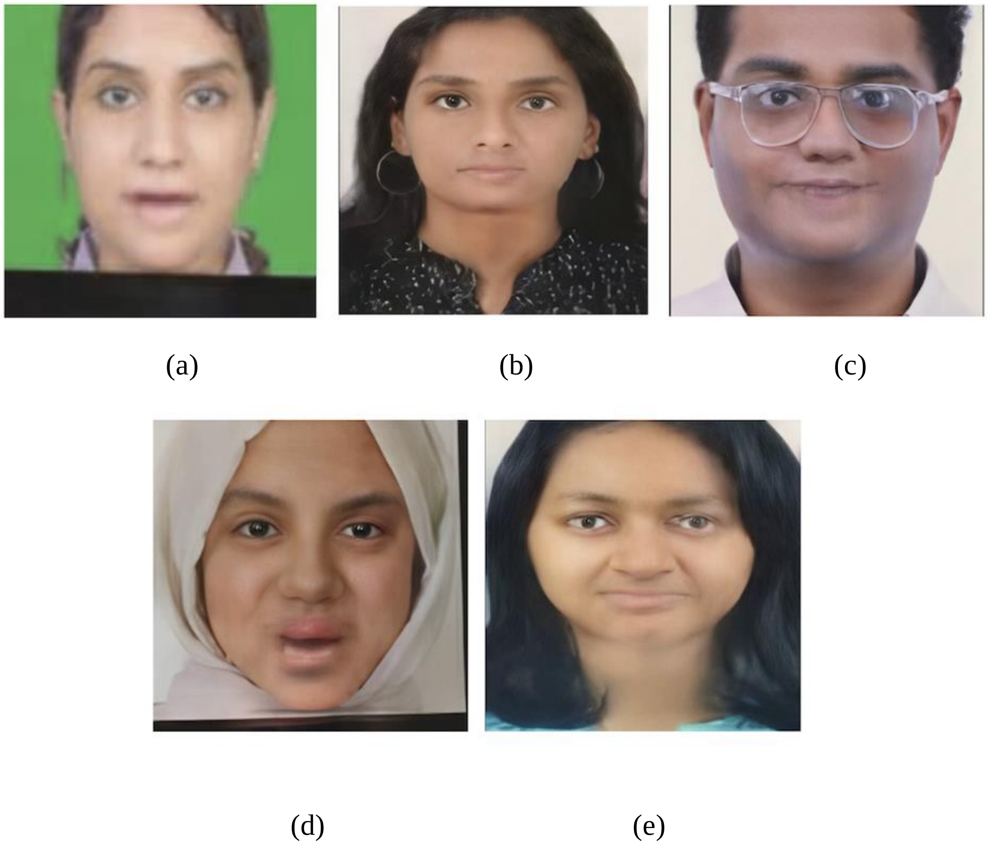
Figure 9: Output of the base paper showing different expressions generated (a) Angry (b) Disgust (c) Fear (d) Happy (e) Neutral (f) Sadness



(a) (b) (c) (d) (e) (f)

Figure 10 Output showing different expressions generated (a) Angry (b) Disgust (c) Fear (d) Happy (e) Neutral (f) Sadness

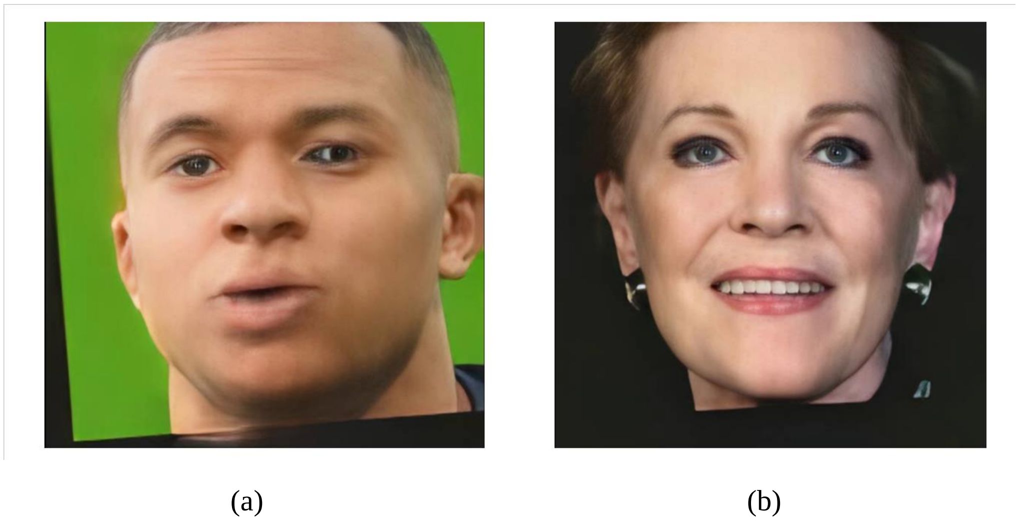
We also attempted to generate outputs using images from the CREMA-D dataset. We used our own face images and celebrity faces like Kylian Mbappé and Julie Andrews to input the model. The results are shown in Figure 11 and Figure 12.



(a) (b) (c)

(d) (e)

Figure 11 (a) (b) (c) (d) (e) Output generated using Group members faces



(a) (b)

Figure 12 Output generated using celebrities' faces

These images showcase the outputs generated by the model using different faces as input.

Overall, the model shows promising results, improved quality when incorporating the SE network, eye mask, and BSRGAN, and the ability to generate facial expressions corresponding to different emotions.

Objective Evaluation

FID

The Fréchet Inception Distance score (FID) is a measure that computes the distance between feature vectors determined for actual and generated images. The score sums up how comparable the two groups are in terms of statistics on computer vision aspects of raw pictures determined using the inception v3 image classification model. Lower scores imply that the two groups of images are more comparable, or that their statistics are more similar, whereas a perfect score of 0.0 indicates that the two groups of images are identical. Lower FID scores have been demonstrated to correlate better with higher quality images when used to evaluate the quality of images created by generative adversarial networks [29].

Table 2: FID Scores

	Objective Evalutation	
Model	FID	% Change
Base	0.001080	-
Base + Squeeze & Excitation (SE)	0.000874	-19.07%
Base + Eye Mask (EM)	0.000753	-30.28%
Base + SE + EM + BSRGAN	0.000142	-86.85%

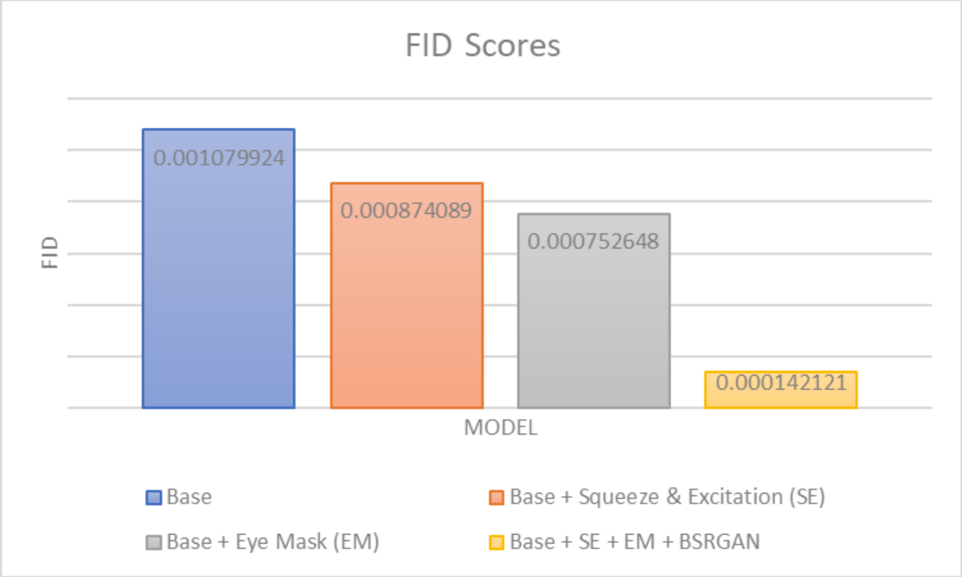


Figure 13: FID Scores

Resolution

Resolution can be an important evaluation metric when we have models generating videos from a single input image. Resolution refers to the level of detail in the generated talking face video, which is determined by the size and quality of the output frames.

Higher resolution can result in more realistic and natural-looking talking face videos, as it enables more fine-grained details to be captured, such as facial expressions, lip movements, and eye movements. In contrast, lower-resolution videos may appear blurry, pixelated, or distorted, which can reduce the overall quality of the generated video. We have improved our base model by adding additional networks, such as BSRGAN and observed that the video's resolution has increased significantly.

Table 3: Resolution

Model	Resolution
Base	128
Base + Squeeze & Excitation (SE)	128
Base + Eye Mask (EM)	128
Base + SE + EM + BSRGAN	512

Bitrate

Bitrate is an important evaluation metric in the context of speech-synchronized talking face generation from an image and an emotional condition. It refers to the amount of data transmitted in a unit of time, and it is commonly used to measure the quality and efficiency of audio and video codecs. Bitrate is commonly measured in bits per second (bit/s).

In the case of talking face generation, the bit rate can affect the quality and clarity of the audio and video components [30]. Higher bit rates can result in better audio and video quality but require more storage space and higher processing power. On the other hand, lower bit rates can result in lower-quality audio and video, but they are more efficient and require less storage space [31]. We have observed that the bit rate of the output video has been improved significantly when we have integrated the additional networks into the base model.

Table 4: Bitrate

Model	Bitrate
Base	107.50
Base + Squeeze & Excitation (SE)	115.00
Base + Eye Mask (EM)	113.33
Base + SE + EM + BSRGAN	3337.83

Subjective Evaluation

MOS

MOS is a subjective assessment metric that is commonly employed in the field of speech synchronised talking face production. MOS's purpose is to create an overall quality metric that is indicative of human perception. Subjective user studies are commonly used to get MOS, in which human evaluators are asked to score the quality of the produced talking face on a Likert scale. A Likert scale is a rating system that is used to assess attitudes or views. A Likert scale, for example, would ask evaluators to grade the created talking face from 1 to 5, with 1 being "very poor" and 5 being "excellent." MOS scores are calculated by averaging all the individual evaluations provided by the evaluators. MOS scores can be useful because they provide an overall assessment of the quality of the generated talking face from the perspective of human observers [32].

Table 5: MOS Scores

Model	How would you rate the motion representation in the video?	How would you rate the video noise level in the video?	Overall how would you rate the quality of the video?	Aggregate (Mean)	% Change from base
Base	2.54	3.13	2.56	2.74	-
Base + BSRGAN	3.44	3.05	2.95	3.15	14.70%
Base + Squeeze & Excitation (SE)	3.18	3.69	3.54	3.47	26.49%
Base + SE + BSRGAN	3.64	4.03	3.82	3.83	39.61%
Base + Eye Mask (EM)	2.83	3.48	2.7	3.00	9.48%
Base + EM + BSRGAN	3.13	3.78	3.04	3.32	20.90%
Base + SE + EM	3.18	3.67	2.96	3.27	19.20%
Base + SE + EM + BSRGAN	3.4	3.88	3.25	3.51	27.95%

The table above shows the scores that the human evaluators had scored for each model. We calculated the aggregate mean and calculated the percentage change w.r.t. the base model.

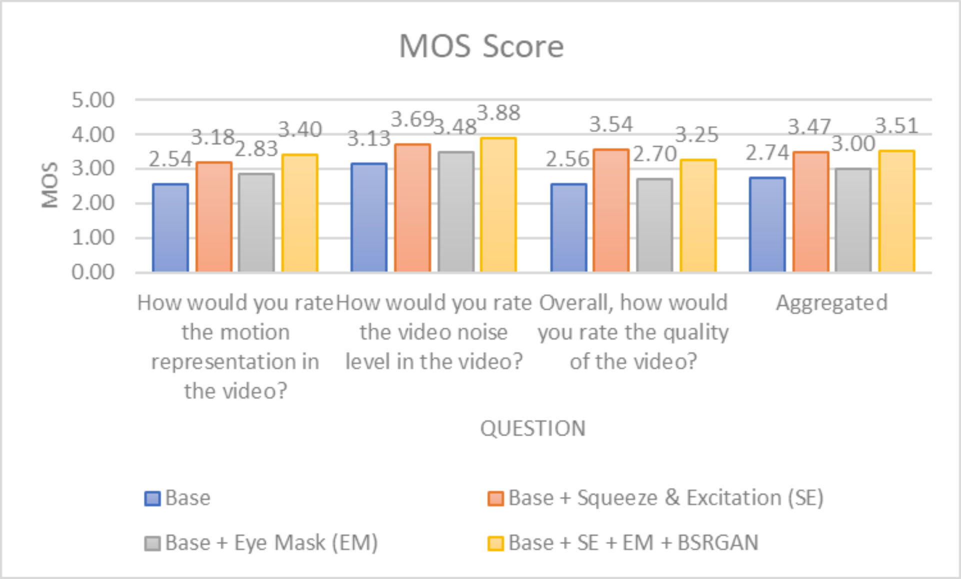


Figure 14: MOS Scores

Table 6: MOS Scores

Model	Do you agree with the following statements?	
	The video was realistic	It was visually Pleasing

	Strongly Disgree (%)	Disagree (%)	Neutral (%)	Agree (%)	Strongly Agree (%)	Strongly Disgree (%)	Disagree (%)	Neutral (%)	Agree (%)	Strongly Agree (%)
Base	20.5	33.3	28.2	15.4	2.6	20.4	38.5	28.2	10.3	2.6
Base + BSRGAN	7.8	17.9	41	28.2	5.1	5.1	28.2	46.2	15.4	5.1
Base + Squeeze & Excitation (SE)	5	10.3	30.8	43.6	10.3	2.4	10.3	38.5	38.5	10.3
Base + SE + BSRGAN	5.1	5.1	35.9	23.1	30.8	2.6	7.7	33.3	33.3	23.1
Base + Eye Mask (EM)	3.5	13.8	31	41.4	10.3	3.6	17.2	37.9	24.1	17.2
Base + EM + BSRGAN	3.5	3.4	27.6	55.2	10.3	3.46	6.9	34.54	37.9	17.2
Base + SE + EM	0.1	10.3	27.6	44.8	17.2	3.4	6.9	34.5	34.5	20.7
Base + SE + EM + BSRGAN	0.1	10.3	10.3	48.3	31	3.5	6.9	27.6	37.9	24.1

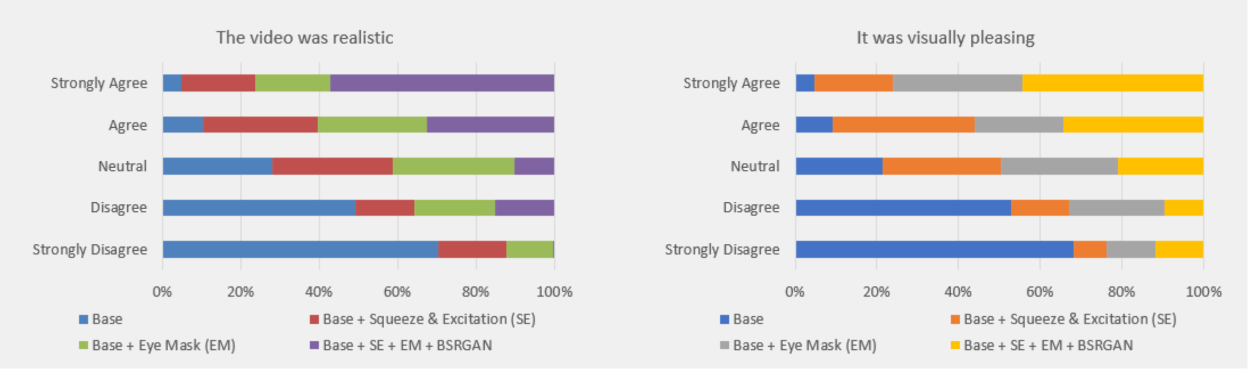


Figure 15: MOS Scores

Conclusion

In this paper, we have successfully implemented and enhanced the performance of a speech driven emotional talking face model that we conditioned on a speech signal, a reference image and categorical emotion inputs. To familiarize ourselves with the working of the model, we initially implemented a GAN model, a simple autoencoder, and an adversarial autoencoder. We had also implemented an Expression GAN model which translates a face image to express a given emotion. Additionally, we performed a literature review of the available literature in the relevant research fields and have outlined our salient findings.

To summarize, this paper presents a significant advancement in the field of speech-driven emotional talking face models, building upon previous work. Through the incorporation of the SE network into the image encoder, the model achieves a notable improvement in performance, giving greater weight to crucial facial features and enhancing the generation of emotionally expressive faces.

BSRGAN, when applied to the output video frames, significantly enhances image quality and resolution, resulting in more visually appealing and realistic talking face outputs. The combination of BSRGAN with other enhancements in our proposed model has led to a substantial increase in the Mean Opinion Score (MOS) by approximately 28.5% from the base model's 2.74 to an impressive 3.51. This demonstrates the model's effectiveness in producing more realistic and visually pleasing talking faces, making the virtual interactions more emotionally engaging.

Additionally, the MTCNN-based eye mask, along with the pre-existing mouth region mask, ensures high-fidelity generation of facial features, further contributing to the model's ability to create lifelike and emotionally expressive talking faces.

The user feedback collected in our study solidifies the remarkable success of the proposed model. An impressive 79.3% of users agreed or strongly agreed that the generated videos were realistic, showcasing a substantial improvement of approximately over the base model's mere 18.0%. Similarly, with regards to visual appeal, 62.0% of users found the proposed model's outputs visually pleasing, marking a remarkable improvement compared to the base model's 12.9%.

Furthermore, the objective evaluation, based on FID scores, supports the efficacy of our proposed model, with a remarkable reduction in FID by approximately 86.9% from 0.001080 in the base model to 0.000142 in our model. This substantial improvement showcases the model's capacity to generate talking faces with higher fidelity, closely resembling ground truth data.

In summary, our speech-driven emotional talking face model, with the integrated enhancements of SE network, BSRGAN, and MTCNN-based eye mask, showcases remarkable performance improvements. The combination of higher MOS scores and reduced FID values signifies the potential of our model in creating emotionally expressive and visually captivating virtual interactions, heralding a new era of immersive human-computer experiences.

Limitations and Future Scope

Limitations

The main limitation we faced was GPU constraints. Retraining the module with a more powerful GPU will help us train our model for the entire dataset and give us possibly better outputs and eliminate current time and GPU constraints.

Future Scope

- The dataset currently does not have largely varied ethnicities for training. Including a diverse set of ethnic faces among others can help improve model outputs for varied images.
- Even though the proposed architecture works on faces that are not from the CREMA-D dataset, there is still a degradation in accuracy compared to faces chosen from the dataset.
- The model can be improved to work on images that are substandard and have low resolution or lighting.
- Further fine-tuning the MTCNN parameters can also give more accurate results with an improved eye loss.

References

[1] N. Rawal and R. M. Stock-Homburg, “Facial emotion expressions in human–robot interaction: a survey,” *International Journal of Social Robotics*, vol. 14, no. 7, pp. 1583–1604, 2022.

[2] J. Tao and T. Tan, “Affective computing: A review,” in *International Conference on Affective computing and intelligent interaction*, Springer, 2005, pp. 981–995.

[3] M. ALPERT, R. L. KURTZBERG, and A. J. FRIEDHOFF, “Transient Voice Changes Associated with Emotional Stimuli,” *Archives of General Psychiatry*, vol. 8, no. 4, pp. 362–365, Apr. 1963, doi: 10.1001/archpsyc.1963.01720100052006.

[4] S. E. 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?,” in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, pp. 2274–2278. doi: 10.1109/ICASSP.2016.7472082.

[5] A. Esposito, “The Perceptual and Cognitive Role of Visual and Auditory Channels in Conveying Emotional Information,” *Cogn Comput*, vol. 1, no. 3, pp. 268–278, Sep. 2009, doi: 10.1007/s12559-009-9017-8.

[6] H. Wang, J. Hu, and W. Deng, “Face feature extraction: a complete review,” *IEEE Access*, vol. 6, pp. 6001–6039, 2017.

[7] P. Ekman and W. V. Friesen, “Facial action coding system,” *Environmental Psychology & Nonverbal Behavior*, 1978.

[8] I. Goodfellow *et al.*, “Generative adversarial networks,” *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[9] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, “Generative Adversarial Networks: An Overview,” *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 53–65, 2018, doi: 10.1109/MSP.2017.2765202.

[10] M. Mirza and S. Osindero, “Conditional Generative Adversarial Nets.” 2014.

[11] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[12] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[13] A. Kammoun, R. Slama, H. Tabia, T. Ouni, and M. Abid, “Generative Adversarial Networks for face generation: A survey,” *ACM Computing Surveys (CSUR)*, 2022.

[14] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[15] J. Xiang and G. Zhu, “Joint face detection and facial expression recognition with MTCNN,” in *2017 4th international conference on information science and control engineering (ICISCE)*, IEEE, 2017, pp. 424–427.

[16] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, “Region attention networks for pose and occlusion robust facial expression recognition,” *IEEE Transactions on Image Processing*, vol. 29, pp. 4057–4069, 2020.

[17] S. E. Eskimez, R. K. Maddox, C. Xu, and Z. Duan, “End-To-End Generation of Talking Faces from Noisy Speech,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 1948–1952. doi: 10.1109/ICASSP40776.2020.9054103.

[18] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[19] X. Wang *et al.*, “ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks.” arXiv, Sep. 17, 2018. doi: 10.48550/arXiv.1809.00219.

[20] K. Zhang, J. Liang, L. Van Gool, and R. Timofte, “Designing a practical degradation model for deep blind image super-resolution,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4791–4800.

[21] L. Song, Z. Lu, R. He, Z. Sun, and T. Tan, “Geometry Guided Adversarial Facial Expression Synthesis.” arXiv, Dec. 10, 2017. Accessed: Aug. 05, 2023. [Online]. Available: http://arxiv.org/abs/1712.03474

[22] Q. Zhang, Z. Liu, B. Guo, D. Terzopoulos, and H.-Y. Shum, “Geometry-driven photorealistic facial expression synthesis,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 1, pp. 48–60, Jan. 2006, doi: 10.1109/TVCG.2006.9.

[23] H. Ding, K. Sricharan, and R. Chellappa, “Exprgan: Facial expression editing with controllable expression intensity,” in *Proceedings of the AAAI conference on artificial intelligence*, 2018.

[24] S. E. Eskimez, R. K. Maddox, C. Xu, and Z. Duan, “Generating talking face landmarks from speech,” in *Latent Variable Analysis and Signal Separation: 14th International Conference, LVA/ICA 2018, Guildford, UK, July 2–5, 2018, Proceedings 14*, Springer, 2018, pp. 372–381.

[25] Z. Fang, Z. Liu, T. Liu, C.-C. Hung, J. Xiao, and G. Feng, “Facial expression GAN for voice-driven face generation,” *The Visual Computer*, pp. 1–14, 2022.

[26] S. E. Eskimez, Y. Zhang, and Z. Duan, “Speech driven talking face generation from a single image and an emotion condition,” *IEEE Transactions on Multimedia*, vol. 24, pp. 3480–3490, 2021.

[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 transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[28] F. Marra, C. Saltori, G. Boato, and L. Verdoliva, “Incremental learning for the detection and classification of GAN-generated images,” in *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, 2019, pp. 1–6. doi: 10.1109/WIFS47025.2019.9035099.

[29] Y. Yu, W. Zhang, and Y. Deng, “Frechet inception distance (fid) for evaluating gans,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3733–3742.

[30] L. Yitong, S. Yun, M. Yinian, L. Jing, L. Qi, and Y. Dacheng, “A study on Quality of Experience for adaptive streaming service,” in *2013 IEEE International Conference on Communications Workshops (ICC)*, Jun. 2013, pp. 682–686. doi: 10.1109/ICCW.2013.6649320.

[31] K. Spiteri, R. Urgaonkar, and R. K. Sitaraman, “BOLA: Near-Optimal Bitrate Adaptation for Online Videos”.

[32] R. C. Streijl, S. Winkler, and D. S. Hands, “Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives,” *Multimedia Tools and Applications*, vol. 51, no. 1, pp. 77–96, 2011.