UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade de Engenharia Elétrica e de Computação

Filipe Antonio de Barros Reis

A Generative Adversarial Network Approach to Visual Expressive Speech Synthesis with Emotion Control

Abordagem por Rede Generativa Adversária para Síntese de Discurso Visual Expressivo com Controle de Emoção

Campinas

2020

Filipe Antonio de Barros Reis

A Generative Adversarial Network Approach to Visual Expressive Speech Synthesis with Emotion Control

Abordagem por Rede Generativa Adversária para Síntese de Discurso Visual Expressivo com Controle de Emoção

Dissertation presented to the School of Electrical Engineering and Computer Engineering of the University of Campinas in partial fulfillment of the requirements for the degree of Master in Electrical Engineering in the area of Computer Engineering.

Dissertação apresentada à Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestre em Engenharia Elétrica, na Área de Engenharia de Computação.

Supervisor: Profa. Dra. Paula Dornhofer Paro Costa

Este exemplar corresponde à versão final da dissertação defendida pelo aluno Filipe Antonio de Barros Reis, e orientada pela Profa. Dra. Paula Dornhofer Paro Costa.

Campinas

2020

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Área de Engenharia e Arquitetura
Rose Meire da Silva - CRB 8/5974

R277g

Reis, Filipe Antonio de Barros, 1990-
A generative adversarial network approach to visual expressive speech synthesis with emotion control / Filipe Antonio de Barros Reis. – Campinas, SP : [s.n.], 2020.

Orientador: Paula Dornhofer Paro Costa.
Dissertação (mestrado) – Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Expressão facial - Simulação por computador. 2. Animação por computador. 3. Computação gráfica. 4. Aprendizado de máquina. I. Costa, Paula Dornhofer Paro, 1978-. II. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Abordagem por rede generativa adversária para síntese de discurso visual expressivo com controle de emoção
Palavras-chave em inglês:
Facial expression - Computer simulation
Computer animation
Computer graphics
Machine learning
Área de concentração: Engenharia de Computação
Titulação: Mestre em Engenharia Elétrica
Banca examinadora:
Paula Dornhofer Paro Costa [Orientador]
Sandra Eliza Fontes de Avila
Gerberth Adín Ramírez Rivera
Data de defesa: 04-08-2020
Programa de Pós-Graduação: Engenharia Elétrica

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0002-3004-4480
- Currículo Lattes do autor: http://lattes.cnpq.br/0823219851164981

COMISSÃO JULGADORA - TESE DE MESTRADO

Candidato: Filipe Antonio de Barros Reis RA: 091202
Data de defensa: 04 de Agosto de 2020
Titulo da Tese: "A Generative Adversarial Network Approach to Visual Expressive Speech Synthesis with Emotion Control"
"Abordagem por Rede Generativa Adversária para Síntese de Discurso Visual Expressivo com Controle de Emoção"

Profa. Dra. Paula Dornhofer Paro Costa (Presidente)
Profa. Dra. Sandra Eliza Fontes de Avila
Prof. Dr. Gerberth Adín Ramírez Rivera

A Ata de Defesa, com as respectivas assinaturas dos membros da Comissão Julgadora, encontra-se no SIGA (Sistema de Fluxo de Dissertação) e na Secretaria de Pós-Graduação da Faculdade de Engenharia Elétrica e de Computação.

Acknowledgements

I acknowledge Prof. Dr. Paula Dornhofer Paro Costa for immense dedication to her students, especially myself. After watching her classes, I have chosen this research topic, and after her careful guidance, I decided on the objective of this work. I feel fortunate to have her as a teacher and tutor. I am also extremely grateful for her patience and extreme care, especially during the review of this work.

I am grateful for the support provided by the colleagues and the companies where I worked during this process, Controllar, Udacity, and Pagar.me.

I want to thank my sister for taking care of me since day one, introducing me to the academic world, and always going out of her way to help me in any possible way.

I want to thank my parents for the unconditional support and precious advice. I am also extremely grateful to have grandparents and uncles to consider as second parents who helped me a lot during the past years.

I wish to thank Larissa Todeschini for supporting me in every possible form. From understanding the extreme dedication I had to put to this work, to giving me the energy to face the challenges and even being a document reviewer whenever I needed.

Abstract

Computer Graphics and Human-Computer Interaction have significantly evolved over the past decade, changing how our society interacts with technology. The interaction with computers and connected electronic devices is shifting from WIMP (Windows, Icons, Menus, Pointer) interfaces to more natural human-like experiences. This shift is heavily related to the advances in speech recognition, text-to-speech synthesis, and natural language processing systems that enabled, for example, the advent of sophisticated virtual assistants that communicate naturally in a variety of situations. However, these assistants still do not have a face. Visual speech communication is naturally multimodal and contains both verbal and non-verbal components. Speech articulatory movements can be modified or modulated by the expression of emotions and other non-verbal communication mechanisms. For this reason, the synthesis of realistic talking-heads and the proper reproduction of facial expressions and speech articulatory movements is a challenging task. This work presents an expressive visual speech synthesis methodology that produces videorealistic results for a talking head's speech. The system adopts a Generative Adversarial Network synthesis approach to produce expressive visual speech using a sequence of facial keypoints as input. The network contains dedicated structures to ensure that the facial expressions match the expressions expected of a given target emotion. To evaluate the work, we analyzed objective metrics and the results of a subjective perceptual study based on the recognition of facial expressions associated with emotions, in addition to a preference test between different synthesis methods. The results demonstrate that our methodology is capable of incorporating facial expressions of a target emotion into visual speech animation, maintaining a high level of videorealism.

Keywords: facial animation, visual speech synthesis, expressive speech animation, machine learning

Resumo

As áreas de computação gráfica e interação humano-computador evoluíram significativamente ao longo da última década, mudando a maneira como nossa sociedade interage com a tecnologia. A interação com computadores e outros dispositivos tem evoluído de interfaces do tipo WIMP (Windows, Icons, Menus, Pointer) para paradigmas mais naturais e similares às interações humanas, tais como a comunicação face-a-face. Essa mudança está muito relacionada aos avanços nas tecnologias de reconhecimento de fala, síntese de texto em fala e processamento de linguagem natural. Tais avanços alavancaram, por exemplo, o surgimento de assistentes virtuais cada vez mais capazes de proporcionarem uma experiência de comunicação natural. No entanto, tais assistentes ainda não possuem uma face. A fala visual é naturalmente multimodal, incluindo componentes verbais (movimentos articulatórios da fala) e não-verbais. Em particular, as expressões não-verbais enriquecem a comunicação e frequentemente influenciam os movimentos articulatórios da fala indicando, por exemplo, se uma frase é uma questão ou afirmação e fornecendo pistas sobre as emoções que acompanham a fala. Pela complexidade das expressões envolvidas na fala acompanhada de expressividade, a criação de cabeças falantes realistas, ou "talking-heads", é uma tarefa desafiadora. Este trabalho apresenta uma metodologia de síntese de animação de fala acompanhada de emoção, resultando numa "talking-head" videorrealista. O sistema utiliza uma rede generativa adversária, do inglês Generative Adversarial Network (GAN), para sintetizar a parte visual da fala com emoção, utilizando como entrada uma sequência de pontos chave da face. A rede contém estruturas dedicadas para garantir que as expressões faciais geradas estejam de acordo com a emoção desejada. Para avaliar o trabalho, foram utilizadas métricas objetivas e resultados de um estudo subjetivo perceptual baseado no reconhecimento de expressões faciais associadas a emoções, além de um teste de preferência entre diferentes métodos de síntese. Os resultados demonstram que nossa metodologia é capaz de incorporar expressões faciais de uma emoção alvo à animação facial, mantendo um alto nível de videorrealismo.

Palavras-chaves: animação facial, síntese da parte visual da fala, animação da fala expressiva, aprendizado de máquina.

List of Figures

List of Tables

List of abbreviations and acronyms

GAN	Generative Adversarial Network
FID	Fréchet Inception Distance
IS	Inception Score

List of symbols

x	Input semantic map
z	Input noise
θ	System parameters
y	Original input data
$\hat{y}$	Synthesized data
P_G	Generator probability distribution
P_x	Real data probability distribution
G	Generator function
D	Discriminator function
e	Target emotion

Contents

1 Introduction

Computer Graphics and Human-Computer Interaction have significantly evolved over the past decade, changing how our society interacts with technology. The interaction with computers and connected electronic devices is shifting from WIMP (Windows, Icons, Menus, Pointer) interfaces to more natural human-like experiences. This shift is heavily related to the advances in speech recognition, text-to-speech synthesis, and natural language processing systems that enabled, for example, the advent of sophisticated virtual assistants that communicate naturally in a variety of situations. This natural communication is enhanced by using speech synthesis systems, allowing users to interact seamlessly with the system.

These virtual assistants may be applied to various applications, such as personal general assistants, where the assistant can execute diverse tasks such as controlling connected devices and retrieving information from the internet. Another interesting application of virtual assistants is the shopping assistants, who are dedicated to helping customers better understand products and make better purchases. Additionally, it is possible to use virtual assistants in healthcare applications, from triage assistants to patient companions, improving the patient experience.

Human speech is naturally multimodal and characterizes an audiovisual signal. The visual cues that accompany the articulatory speech movements are part of the input information processed by the brain to extract meaning from the speech. If we talk to someone in a noisy environment, we naturally switch to a vision-based speech intelligibility processing. An enhanced visual speech processing mechanism is also developed by individuals that are deaf or hard of hearing.

Visual information on the face also adds key elements of non-verbal communication to the speech, which are essential to social interactions. These elements express (involuntarily or in a controlled manner) the informant's internal states such as emotions (facial expressions that accompany speech), tiredness (for example, through excessive blinking of the eyes), and intentions (pointing the look at some direction or complementing the speech with a nod) (MATTHEYSES; VERHELST, 2015).

Considering the role of visual information in speech, adding embodiment to virtual assistants is vital to improving communication and promoting more natural, engaging, and intuitive interaction mechanisms. Studies about synthetic audiovisual speech have demonstrated that the addition of visual speech signal leads to more positive reactions (PANDZIC I., 1985) and better engagement with the agent (WALKER; SPROULL; SUBRAMANI, 1994). Kim et al. (2019) present a study on the effects of patient care

assistant embodiment, concluding that this embodiment resulted in a more engaging interaction. Fraser, Papaioannou e Lemon (2018) show that the modeling of character emotions on a role-playing video game, helped to turn the system more engaging, making the users spend more time talking to the game agent and improving the evaluation of users about the system, turning it more enjoyable to use.

In particular, the expression of emotions by embodied conversational agents provides the ability to demonstrate empathy to the user, which is extremely valuable in customer-facing applications, such as educational applications and support centers. An agent capable of expressing emotions can react much better to the user's emotions and interact accordingly. However, the synthesis of expressive agents is still an open challenge due to its high complexity. Incorporating emotions on an agent dramatically increases the synthesis effort once the speech needs to be carefully synthesized for each emotion available to the agent, and the facial expressions need to be adapted accordingly.

Considering this vital role that visual speech has in improving communication, the present work adopts an image-based, or 2D, approach to synthesize expressive talking heads.

The typical approach to 2D facial animation synthesis has been using data-driven models, either by reusing real pieces of the original speech or using statistical models trained on the original speech (MATTHEYSES; VERHELST, 2015). The statistical model approach is commonly based on Hidden Markov Models (HMM).

Advances in easily available computational power have also allowed significant advances in the fields of Neural Networks. One of such advances was the creation of Generative Adversarial Networks (GOODFELLOW et al., 2014), which allows the computer to create a model capable of generating new data based on a training set. This family of networks has been typically used to create new images on applications such as face swapping and image remapping (HUANG; YU; WANG, 2018).

We propose the use of Generative Adversarial Networks to fully synthesize images from keypoints, obtaining a data-driven system trained on a set of videos capable of synthesizing 2D videorealistic expressive visual speech for an agent face using only facial keypoints coordinates as input. This approach allows for generating different faces, requiring only the retraining of the model. The possibility of synthesizing different faces with the same input keypoint is extremely important as it allows the customization of virtual assistants, and even face transfer.

1.1 Objectives

The widespread acceptance and increase in the use of virtual assistants with visual representations are only possible if they are natural, accurate, and have good quality. This project has the general objective of implementing and evaluating a talking head synthesis system using Generative Adversarial Networks.

Our specific objectives in this project are:

1. To investigate the state of the start of expressive facial animation synthesis;
2. To propose an expressive visual speech synthesis pipeline using Generative Adversarial Networks;
3. To propose and to conduct an evaluation protocol to assess the results.

1.2 Contributions

The main contributions of this work are:

- *Creation of a system capable of generating the visual part of the speech of a talking head while expressing emotions*: we create a system capable of generating expressive visual speech based on only a sequence of facial keypoints and a target emotion. Additionally, this system is capable of achieving better results than the state-of-art *vid2vid* network. Finally, the proposed system can generalize well to keypoints extracted from new faces;
- *Creation of an evaluation protocol for emotion perception and video quality assessment*: we propose a complete evaluation protocol used to obtain our results. Our evaluation system can perform tests for comparing the quality between distinct videos and for defining which emotion the users can perceive on a given video. the whole process is randomized to avoid positional and sequential biases;
- *Evaluation of the effects of adding emotion to talking head video*: we evaluate the influence that producing expressive visual speech has on the perception of the videorealism of the speech.

The contributions of this work also resulted in the following publication:

- Filipe Antonio de Barros Reis, Paula Dornhofer Paro Costa, and José Mario de Martino. 2020. Deeply Emotional Talking Head: A Generative Adversarial Network Approach to Expressive Speech Synthesis with Emotion Control. In Special Interest

Group on Computer Graphics and Interactive Techniques Conference Posters (SIGGRAPH '20 Posters), August 17, 2020, Virtual Event, USA. ACM, New York, NY, USA 2 Pages. https://doi.org/10.1145/3388770.3407417

1.3 Text Organization

This text is organized as follows. The core concepts of Generative Neural Networks are presented in Chapter 2, as well as the related work to this research. In Chapter 3, we present our methodology and objective results using this method. We present the perceptual study considered in the evaluation process, as well as the results in Chapter 4. Finally, we present final remarks and directions for future work in Chapter 5.

2 Background

This chapter discusses the state of the art of expressive facial animation focusing on the use of Generative Adversarial Networks. The first section of this chapter provides an introduction to Generative Adversarial Networks and how they work. Section 2.2 presents a review of visual speech synthesis and state of the art for expressive facial animation. In Section 2.3, we discuss the role of an emotion model and review the use of these models in the state of the art of expressive visual speech synthesis. The final section of this chapter provides a discussion on the relation of the current work with the existing approaches.

2.1 Generative Adversarial Networks

Generative Adversarial Networks (GANs) are a specific set of neural network arrangements capable of mapping a given distribution and generating new data with similar characteristics. After properly trained, these networks generate new, realistic data following the same distribution of the original data. The generated data may be in different formats, such as image, audio, and temporal series. This generation process is useful in many applications, such as generating new data to train semi-supervised or reinforced models (GOODFELLOW, 2017). In this work, we focus on applying GANs to generate images, as by generating a sequence of images of faces, we can obtain expressive speech animation.

The first Generative Adversarial network was proposed by Goodfellow et al. (2014) as a framework for estimating generative models. Although there have been significant advances in GANs, the core concepts presented in this section are still used (GOODFELLOW, 2017).

GANs may be considered as an arrangement of neural networks because they are composed of a combination of two distinct networks. The networks that compose a GAN are the *generator* and the *discriminator*. The generator is the network responsible for capturing a given data distribution well enough to be capable of generating new data following such distribution. The discriminator is a network responsible for detecting real and fake data, being the real data a sample from the dataset, and the fake synthesized by the generator. In Figure 2.1, we graphically present this relation representing inputs and outputs as parallelograms, networks as rectangles and decisions as diamonds. Discriminator structures are presented in green and generator in blue, while results produced by the system are yellow, and external data is purple.

The adversarial part of this network arrangement relates to the training phase,

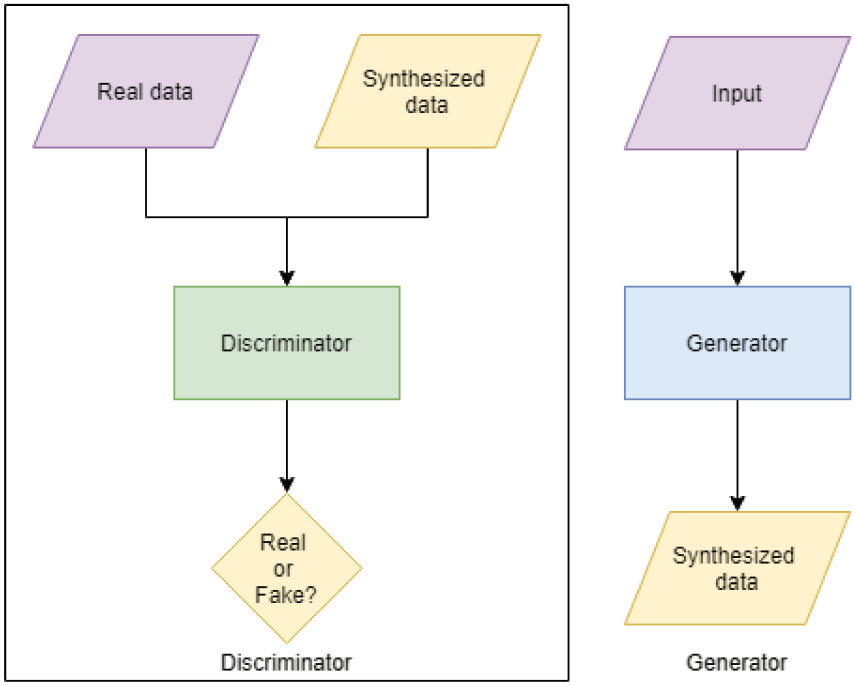


Figure 2.1 – A Generative Adversarial Network is composed of two neural networks, the discriminator, in green, and the generator, in blue. The discriminator is responsible for determining if its inputs are real (purple) or produced by the generator (yellow). The generator is responsible for generating new data using a given input data.

when the generator and the discriminator compete, allowing the generator to learn the distribution of the data. This competition occurs because the generator aims to synthesize images classified as real by the discriminator. In contrast, the discriminator tries to classify which data is real and which is synthesized correctly.

We may define the generator mathematically as the function $G(x;\theta_g)$, where G is the function associated with the generator, θ_g the respective parameters and x is the input for the function, which may be an input noise. For simplicity, we omit θ_g in our representations. P_g is the probability distribution function defined by the generator function G. The goal of the generator is to make P_g as close as possible to the probability distribution function of a given real dataset, represented by P_y, where y is the set of the real data. Similarly, the discriminator may be represented by $D(y;\theta_d)$, where D is a function with parameters θ_d, and $D(y)$ represents the probability that y came from the real data instead of P_g. Due to the use of backpropagation during the training of a GAN, both the discriminator and the generator functions need to be differentiable.

The training stage of a GAN may be divided into four steps, as shown in Figure 2.2. The first step is the generation of new data by the generator given a single input. The second step is composed of the discriminator analyzing the data that the generator has just synthesized, along with samples from the training dataset, to evaluate which are real and fake. After this, a loss is calculated for the generator and the discriminator. The

loss for the generator penalizes instances when the discriminator was able to classify the synthesized data as fake correctly. In contrast, the loss for the discriminator penalizes when it wrongly classifies the original sample or the synthesized data. The last step of the training process is the backpropagation of the losses to update the weights of the networks.

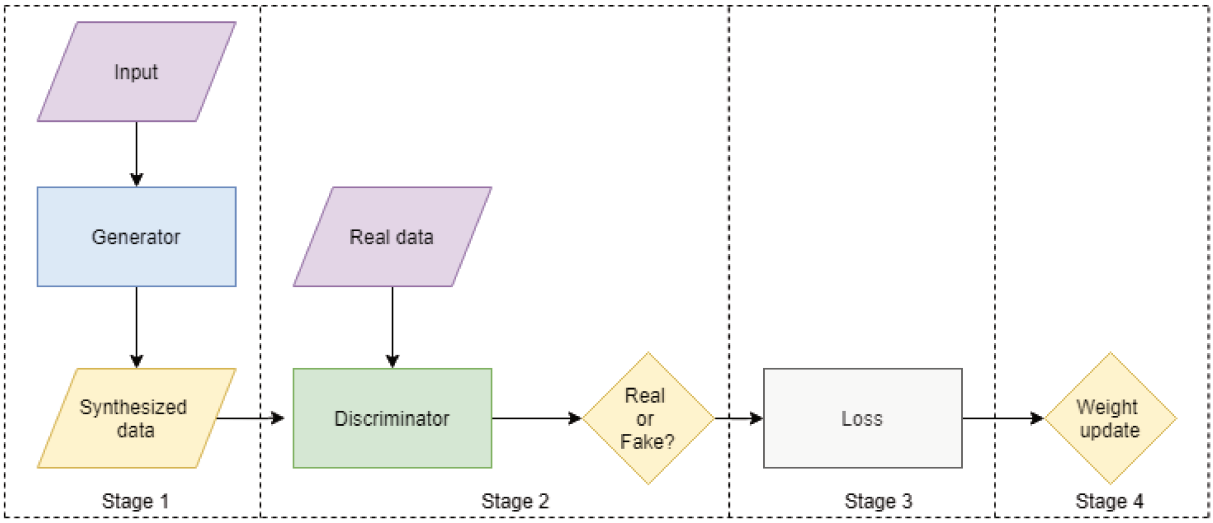


Figure 2.2 – Training process of a simple GAN. In the first stage of training, the generator uses some input to synthesize new data. During the second stage, the discriminator analyzes the synthesized data and samples of real data to determine which are real and fake. Then the loss is calculated by penalizing the discriminator for miss classification of inputs and the generator for when the discriminator identified the synthesized data as fake. After the loss calculation, the weights for both the discriminator and generator networks are updated through a backpropagation step

On the original work by Goodfellow et al. (2014) and many of its successors, the adopted loss function was the binary cross-entropy loss, as defined on Equation 2.1, where y is the original data and $\hat{y}$ is the reconstructed data. This loss is used because we want to assure that the probability distribution mapped by the generator, P_g, is as close as possible to the one associated with the real samples, P_y. We use the binary loss because there are only two possible classes in this context, real and fake.

$$\mathcal{L}(\hat{y}, y) = [y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})]. \tag{2.1}$$

Applying to the GAN context, during the discriminator training, we define that y is 1 for real data. At the same time, $\hat{y}$ is the result of the discriminator for the input x, $D(x)$. Applying this to Equation 2.1 yields in Equation 2.2, which is valid for classifying real data.

$$\mathcal{L}(D(y), 1) = \log(D(x)). \tag{2.2}$$

When the discriminator is classifying fake data, we define that y is 0, while $\hat{y}$ is $D(G(x))$ as now the discriminator is analyzing the result from the generator. This yields in Equation 2.3, used for the discriminator classification of fake images.

$$\mathcal{L}(D(y), 0) = \log(1 - D(G(x))). \tag{2.3}$$

As the goal for the discriminator is to classify the data correctly, the objective of this network is to maximize Equation 2.4, which is the equivalent to perform the operation shown in Figure 2.3.

$$\mathcal{L}(D) = \max[\log(D(y)) + \log(1 - D(G(x)))]. \tag{2.4}$$

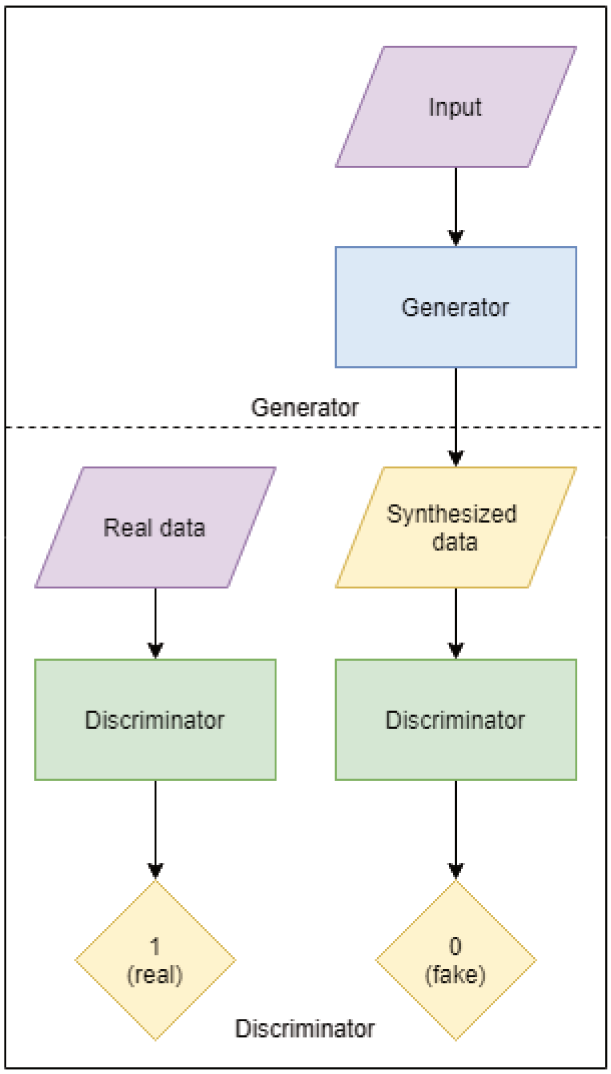


Figure 2.3 – The goal for the discriminator of the Generative Adversarial Network is to correctly classify real and fake samples. We consider a discriminator output of 1 as a real sample and 0 as a fake one. The objective during training is to maximize the likeliness of the discriminator correctly classifying the samples.

As the generator and the discriminator compete, the objective of the generator is to minimize the correct classification by the discriminator as presented on Equation 2.5, which is the equivalent to perform the operation shown in Figure 2.3. This means that the generator aims to produce synthetic results indistinguishable from real ones.

$$\mathcal{L}(G) = \min[\log(D(y)) + \log(1 - D(G(x)))]. \tag{2.5}$$

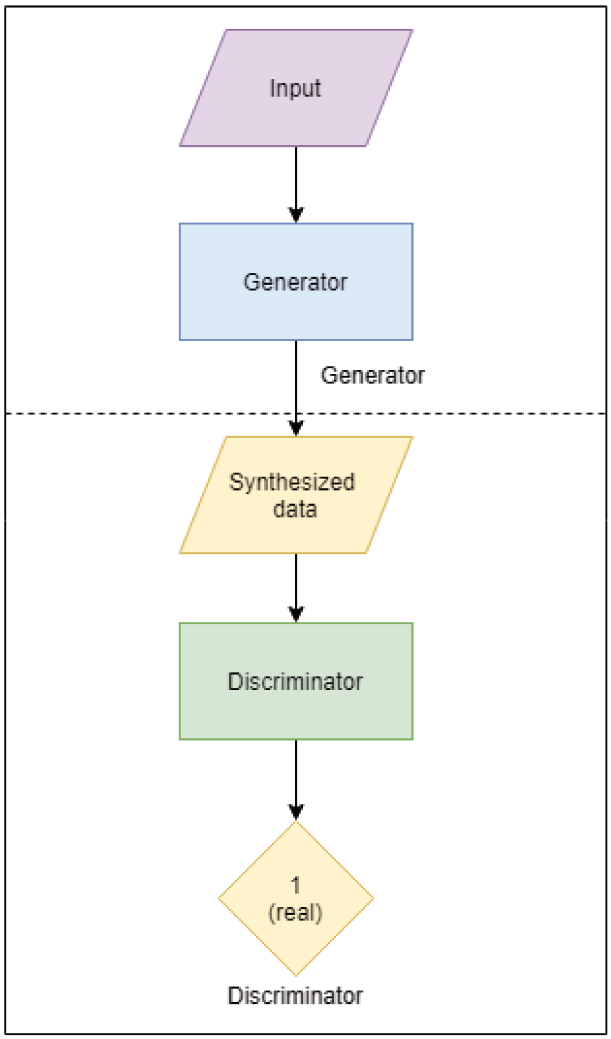


Figure 2.4 – The goal for the generator of the Generative Adversarial Network is to have its synthesized samples classified as real by the discriminator. In our notation, this translates into an output of 1 to the discriminator.

Considering both networks as part of the GAN, the general loss function is the combination of Equation 2.4 and Equation 2.5, which results in Equation 2.6:

$$\mathcal{L}_{GAN} = \min_{G}\max_{D}[\log(D(y)) + \log(1 - D(G(x)))], \tag{2.6}$$

where:

- $\mathcal{L}_{GAN}$ is the general loss function for the network;

- $D(G(x))$ is the result for the probability estimation that a synthesized instance is real;
- $D(y)$ is the result for the probability estimation that a real data instance is in fact real.

After training, new samples are generated using the sequence shown in Figure 2.5, where the trained generator uses the input to generate fake images with a distribution similar to the one from the original set.

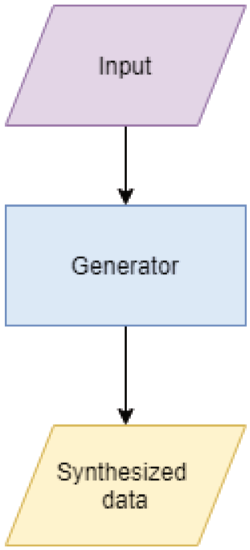


Figure 2.5 – After the training process is complete, the generator can synthesize new samples similar to real ones from a single input, as its parameters were optimized during training to fool the discriminator into considering the synthesized samples as real.

The functions for the discriminator and generator need to be differentiable, as the training process depends on backpropagation for the calculation of weights. In the first approach to GANs, Goodfellow et al. (2014) used single multilayer perceptrons for both the discriminator and generator.

Using random noise as input may allow good results, but there is no control over the output. To add this output control, Mirza e Osindero (2014) introduced the use of segmentation maps as input. Each image from the training set is paired with a corresponding segmentation map, which allows the network to capture the relation between the segmentation maps and the outputs, generating results that are realistic and aligned with the maps. Equation 2.7 shows the objective function for a Conditional GAN:

$$\min_G \max_D V(D,G) = \mathbb{E}_x[\log(D(x|y))] + \mathbb{E}_z[\log(1 - D(G(x)))], \tag{2.7}$$

where:

- x are the inputs to the generator, the segmentation maps corresponding to the real inputs y;
- $\mathbb{E}_x$ represents the expected value over the real data instances;
- $\mathbb{E}_z$ represents the expected value over the synthesized instances;
- $V(D, G)$ represents the value function for the two-player min-max game between the generator and the discriminator.

Conditional GANs are a significant increment over traditional networks, especially for face generation. It is possible to use as input facial keypoints and obtain a resultant model capable of generating visual speech using only new facial keypoints.

To further improve the results, Radford, Metz e Chintala (2015) proposed the use of Convolutional Neural Networks instead of Multilayer Perceptrons as networks for both the generator and the discriminator. These networks provide a significant performance increase as the structure of CNNs is better suited for images.

In addition to changing the generator and discriminator functions, researchers have developed other alternatives to improve GANs, such as changing the loss function and using regularization techniques to prevent mode collapse (HONG et al., 2019).

2.1.1 Objective Metrics

Evaluating GANs that produce images and videos is challenging, as the quality of these results is subjective and closely related to human evaluation. Another factor that increases the difficulty is that the loss calculated during training is a bad indicator of actual performance compared to what happens on classification problems. This difference between the training loss and the actual performance of a GAN occurs because a combination between the generator and discriminator losses composes the general loss of a GAN. As these losses are determined in an adversarial manner, it is possible to achieve good general loss when one network dominates the other, which results in the bad visual output. As the studies with GANs move fast and with a big scale of results, researchers have adopted other objective measures to help.

These standard objective metrics usually aim to determine the level of realism of the generated data. One standard objective metric is the *Inception Score (IS)*, which focuses on evaluating the quality and diversity of the results produced by the GAN. *IS* uses the *Inception-v3* neural network created by Szegedy et al. (2016) to classify the results of the GAN. This process generates a set of class probabilities for each of the 1000 classes in this network and each result. It is then possible to determine if the image contains meaningful objects by analyzing if the conditional class probabilities have low

entropy (SALIMANS et al., 2016). Additionally, to identify if the results are varied, the *IS* analyzes if the integral of the marginal probability distribution has high entropy. *IS* calculates the average relative entropy between conditional and marginal probabilities, which is the final output of this metric (SALIMANS et al., 2016). The best score possible for *IS* when trained using the *InceptionV3* model trained on the ILSVRC 2012 dataset is 1000, as this is the number of classes in the dataset. In contrast, the lowest score is 1. One major drawback of the *IS* is that this metric does not consider any information on real samples (HEUSEL et al., 2017).

To improve the drawback of *Inception Score*, Heusel et al. (2017) proposed the *Frechet Inception Distance* (*FID*). Recent studies, such as (ZAKHAROV et al., 2019; WANG et al., 2018; WANG et al., 2019), vastly use this metric due to its ability to capture the similarity between real images and synthesized ones. *FID* also uses results from the *Inception-v3* network to capture features from the images for both real and synthesized data. These features are then summarized into a probability distribution, and the distance between the distribution of real and fake images is calculated using the *Frechet Distance*. This process makes the *FID* more robust and with results closer to the human evaluation on the realism and variability of the results (HEUSEL et al., 2017). Smaller *FID* results indicate a more realistic dataset than bigger ones. One drawback of *FID* is its inability to detect overfitting, so if the considered network only reproduces real images, it will achieve a good score (LUčIć et al., 2018). Another shortcoming of *FID* is that the *Inception-v3* is typically trained on a dataset with many distinct objects, which makes the metric less precise for comparing results between items from a single class, like a human face (LUčIć et al., 2018).

Another group of objective metrics is formed by the techniques to measure the similarity of the synthesized output with the original, ground-truth, video. These metrics evaluate each frame and calculate the distance from the original to the synthesized one. Many works on facial animation and expressive speech synthesis use this group of metrics to evaluate their results, especially the *Structural Similarity (SSIM)* metric (MATTHEYSES; VERHELST, 2015). *SSIM* and other video similarity techniques are less used when designing GANs as these metrics measure the ability of the results to replicate the original images accurately. This objective is not always aligned with the goal of these networks, as they aim to create entirely new images.

In contrast, these metrics may be applied to our case as we consider only one actress in our training, so comparing the results obtained using the same keypoints from the training dataset with the real video indicates the quality of the results. Additionally, temporal consistency is essential for visual speech synthesis, and both *FID* and *IS* consider the results as a set without a given order, while *SSIM* accounts for the temporal consistency. Finally, *SSIM* allows a comparison between our results and others from previous works.

2.2 Related Work in Facial Animation Synthesis

The creation of automated videorealistic talking heads has been a research topic since the 1972 pioneering work, from Parke (1972). As described by Mattheyses e Verhelst (2015), researches in this area have adopted many different synthesis strategies. In the present work, we focus on the strategy adopted to generate facial dynamics during the speech. Besides, the target of the synthesis can be in either 2 (image-based) or three dimensions (model-based). In this section, we consider only approaches that generate image-based results.

The first strategy is a rule-based synthesis, where a set of predefined rules are applied to determine the movement of a head model for each synthesized frame. The frames that are not synthesized originate from a dataset, allowing these systems to be also called *keyframe-based*. Rule-based systems have proven to be quite versatile, allowing the creation of a wide range of expressive agents in 2D, such as Liu e Ostermann (2011), Ezzat, Geiger e Poggio (2002). In Costa (2015), the authors propose a rule-based system capable of synthesizing expressive video. This strategy relies on rules for the transition between visemes. It requires a set of annotated visemes that covers as many transitions as possible, significantly increasing the cost associated with the creation of the dataset.

Concatenative systems can create new speech data by using segments that match the required phoneme sequence, forming a new set of segments, that are then concatenated. These systems have the advantage of requiring a dataset only as big as the number of possible combinations of diphones. A pioneer work to use the concatenative method is *Videorewrite* (BREGLER; COVELL; SLANEY, 1997), where existing images from an actor speaking are used to generate new speeches. This is achieved by either evaluating the phoneme sequence and reordering the existing frames or interpolating similar frames when the desired one is not initially available (MATTHEYSES; VERHELST, 2015). Another approach using this technique was presented by Cosatto e Graf (2000), when the authors proposed a system with improved cost function to determine the best sequence of original videos from the dataset for the synthesized one. A significant advantage of Mattheyses e Verhelst (2015) is that the authors used videos acquired without controlled conditions such as front-facing and TV studio environment. A downside of this technique is that as it relies on existing frames, there are only a few concatenative expressive systems.

The last approach, model-based synthesis, rely on mathematical models to generate new speech. Researches using this approach can obtain a model by exposing a training set to either a statistical or, more recently, a Neural Network model. Many works that use Hidden Markov Models as the driver of the synthesis, such as (ANDERSON et al., 2013; XIE; SUN; FAN, 2014; FAN et al., 2015). Recently, there is a shift from the HMM models to models based on Deep Neural Networks, especially Generative Adversarial Networks.

Generative Adversarial Networks were proposed by Goodfellow et al. (2014) as a framework for estimating generative models, which was able to generate new images, with a distribution similar to the present on a data set, using random noise as input. Further studies improved the versatility of GANs by changing the input from random noise to segmentation maps, such as facial keypoints or contour lines (MIRZA; OSINDERO, 2014). Another vital contribution to GANs was the adoption of Convolution Neural Networks as the internal neural networks instead of Multilayer Perceptrons. This change provided a significant performance improvement for images (RADFORD; METZ; CHINTALA, 2015).

Another vital contribution to GAN structure was the use of patch GANs, which improved the efficiency while maintaining the performance of the discriminator by using a sequence of convolutions layers and analyzing smaller patches (LI; WAND, 2016). In Vondrick, Pirsiavash e Torralba (2016), the authors propose a GAN for video *VGAN*, using a Spatio-temporal convolutional architecture that learns typical video dynamics from a massive set of unlabeled videos.

The training process of GANs is extremely challenging due to its adversarial nature, which leads researchers to propose new optimization strategies, such as using special activation functions different architectures, such as U-Net (RONNEBERGER; FISCHER; BROX, 2015). To reduce the necessity of having a paired dataset, with both original images and their segmentation maps, Zhu et al. (2017) proposes a cyclic loss strategy, which adds an extra step of transformation, transforming a segmentation map to a target image and from target image back to segmentation map, to train based on the consistency of the transformations.

With these improvements, GANs started to generate even better results on various applications, synthesis of faces, and expressive speech. In *pix2pix*, both the source and target domains are images, and a key element to the success of this network is that it learns not only the mapping but also a loss function to train this mapping, which excludes the need to hand engineer features (ISOLA et al., 2017). Although *pix2pix* provides excellent performance to static images, the performance for videos is not optimal, as the frames are synthesized sequentially, without the knowledge of prior results, degrading continuity. To improve this performance when generating videos, *vid2vid* proposes the use of a network structure similar to *pix2pix*, but with an additional discriminator for analyzing the images generating when computing the optical flow of the previous images(WANG et al., 2018). By doing so, *vid2vid* differs from *VGAN* because it calculates the next frame by analyzing the relation between the previous ones. In contrast, the latter calculates the next frame based on a prediction learned from other videos. Although *vid2vid* provides videorealistic results, the output is a direct mapping from the input, making the task of conveying emotions difficult, as there is no control of the outputs concerning the target emotion.

Recently, Zakharov et al. (2019) used a Conditional GAN in combination with pre-trained models to employ meta-learning in the process and achieve great results with as few as one input image of the target. In a similar line of work, (WANG et al., 2019) uses attention models to achieve videorealistic results with only a few input frames, as little as two frames. Both these works offer only the mapping from keypoints to images, without any possibility of controlling additional aspects such as emotions.

An important line of work presented in Karras, Laine e Aila (2019), Karras et al. (2020) uses concepts from style transfer to perform unsupervised separation of attributes such as pose and identity. These approaches can produce photorealistic results without requiring semantic maps as inputs.

Another use of Generative Adversarial Networks for the synthesis of a talking-head is using GANs to map sketches of faces to actual images (Kazemi et al., 2018; HU; GUO, 2020). Although this approach provides a possible way of generating a talking head, the difficulty of obtaining the sketches makes it not optimal for this use. Additionally, the results are not videorealistic, as the frames are generated without considering the previous outputs.

Another line of work related to facial synthesis is facial editing. Recent ground-breaking work is proposed by Suwajanakorn, Seitz e Kemelmacher-Shlizerman (2017), where facial dubbing was possible by using a Recurrent Neural Network and a large dataset containing more than 7 hours of video. Although the results were expressive, the amount of input necessary to train was too extensive. In Fried et al. (2019), the authors combine Recurrent Neural Networks with a GAN to allow video editing based on phonemes as input, achieving great results for making corrections after a video was filmed.

Finally, some works use GANs for synthesizing still images with new facial expressions. Aiming to generate more inputs to be used on emotion recognition tasks, Wang et al. (2019) proposes a method to synthesize new emotions on still image, using an image of a face as input, and generating the same face but expressing a different emotion as a result. As the goal of this work was to improve facial expression recognition performance, no perceptual evaluation was performed. Qiao et al. (2018) proposes a Geometry-Contrastive GAN method to transfer emotions across different subjects, and handle emotion transition. However, the evaluation of this work is based on similarity metrics rather than user recognition of emotions. Our approach is capable of generating different emotions with only the base facial keypoints, without requiring a real picture as input during the synthesis of new frames. These methods focus on still images, which makes the results not videorealistic, as they are generated from pure segmentation mapping and do not consider the previous frames on the generation process.

As our focus is to generate videorealistic results capable of conveying emotions with control, our work expands the structure of *vid2vid* by adding the ability to achieve

expressive visual speech capable of conveying the desired emotion while maintaining realism and time consistency. As described in Chapter 3, we add the target emotion information on the segmentation maps and include one emotion discriminator to guarantee proper system training. To generate the segmentation maps, we use a sequence of keypoints to determine the desired output shape, allowing the system to render the new face with this desired shape.

2.3 Models of Emotions

The use of emotions on facial animation synthesis is a challenging task, as the expression of emotions is naturally multimodal and complex because the emotion conveyed is a combination of the positioning and movement of different aspects of the face, such as eyebrows, lips, nose, and forehead.

Another challenge creating systems capable of synthesizing facial animation with emotions is that the dataset required to train the system is significantly bigger. Usually, we need the same amount of samples for each target emotion, which escalates the number of total samples needed (MATTHEYSES; VERHELST, 2015). For systems that require at least one sentence with each context-dependent viseme of the language (MARTINO; MAGALHãES; VIOLARO, 2006), adding emotions greatly increases the required dataset size.

Most recent work on expressive facial animation using neural networks is based on producing results with a small set of categorical emotions. This is due to the availability of datasets containing these emotions.

In Qiao et al. (2018), researchers aim to transfer the emotion of a face to another by using a Geometry-Contrastive GAN. To validate the proposed network, the authors use different datasets with discrete categorical emotions and produce quantitative results. Comparing the presented video similarity metrics, PSNR and SSIM, this work achieves good results, which means that the final image is similar to the original one, which is related to the objective of maintaining the original face characteristics. In contrast, the researchers did not perform any emotion recognition tests, not allowing them to verify the accuracy regarding the emotions synthesized.

Another approach to emotion transfer using GAN is presented by Choi et al. (2017) as an application to the proposed unified GAN for multi-domain image-to-image transfer. The authors used the RaFD dataset, which contains images of models expressing eight still emotions, to train a model capable of synthesizing new images expressing different emotions for a single image of a neutral face. This work also considers only quantitative metrics and does not present the performance of the model on the emotion recognition task.

Although most recent works consider a small categorical set of emotions, the context in which talking heads may be used probably requires more complex emotions. Costa (2015) proposed the adoption of the Ortony, Clore, and Collins (OCC) model of emotions for embodied conversational agents, as these emotions are adequate for the application. This model is well suited for synthesis tasks as it contains 22 different subtle emotions defined through a valenced reaction to different events, actions, or aspects. The present work is a sequence to (COSTA, 2015), and we use the dataset proposed in that work, which contains carefully designed sentences for each emotion in the OCC model.

2.4 Concluding Remarks

This chapter presented an introduction of Generative Adversarial Networks (Section 2.1), which are the foundation of the system proposed in this work. It also presented a review of related work in facial animation synthesis, focusing on projects using GANs in Section 2.2.

In Section 2.3, we have discussed the use of emotion models on relevant work and the importance of using such models.

The following chapters present a Generative Adversarial Network approach to expressive facial animation, which produces a talking head capable of conveying the target emotions. To validate that the synthesized samples convey the desired emotions, we perform a subjective test with a user group to evaluate if the users can discern the emotions presented by the talking head.

3 Expressive Visual Speech Synthesis

Recent advances in GANs have allowed visual speech synthesis with great videorealism and overall quality. These networks can synthesize realistic video speech, as described in Section 2.2. Despite the quality, these networks do not focus on conveying emotions in a controllable way. In this chapter, we propose a network capable of synthesizing realistic video speech while expressing a defined emotion.

We adopted the *vid2vid* network as the foundation for our approach. We discuss how this network works in Section 3.1. Following, we discuss our key contributions to improve the results of *vid2vid*: the emotion discriminator in Section 3.2.2 and the emotion segmentation maps in Section 3.2.1. Section 3.4 presents the experiments we performed during the development of our network to validate our proposal. Figure 3.1 illustrates the components used in our network, along with the section where we describe each component.

3.1 Vid2vid Network

The *vid2vid* network is a GAN composed of multiple discriminators and generators to obtain good results on image quality and video realism. The authors have demonstrated the possibility of using this network for various domain transformation applications, including facial keypoints to facial synthesis (WANG et al., 2018).

In Chapter 2, we presented the concept of a conditional GAN. The *vid2vid* network uses the conditional GAN concept as a foundation, but with a more complex structure. In this section, we describe each component of the *vid2vid*, and in Section 3.2 we present our proposed network.

3.1.1 Generator

The *vid2vid* network aims to transform sequences of edge or semantic segmentation maps into realistic video frames. This differs from a standard conditional GAN as we need to generate a sequence of frames to form a video. We may model this relation by considering $y_1^K = \{y_1, y_2, y_3, ..., y_K\}$ as the sequence of real frames, and $x_1^K = \{x_1, x_2, x_3, ..., x_K\}$ as the corresponding semantic segmentation maps, where K is the total number of frames. We define $\hat{y}_1^K = \{\hat{y}_1, \hat{y}_2, \hat{y}_3, ..., \hat{y}_K\}$ as the sequence of output frames. Equation 3.1 shows the main objective of the generator, which is to produce results with a distribution similar to the original distribution of the training data.

$$P(\hat{y}_1^K | x_1^K) = P(y_1^K | x_1^K), \tag{3.1}$$

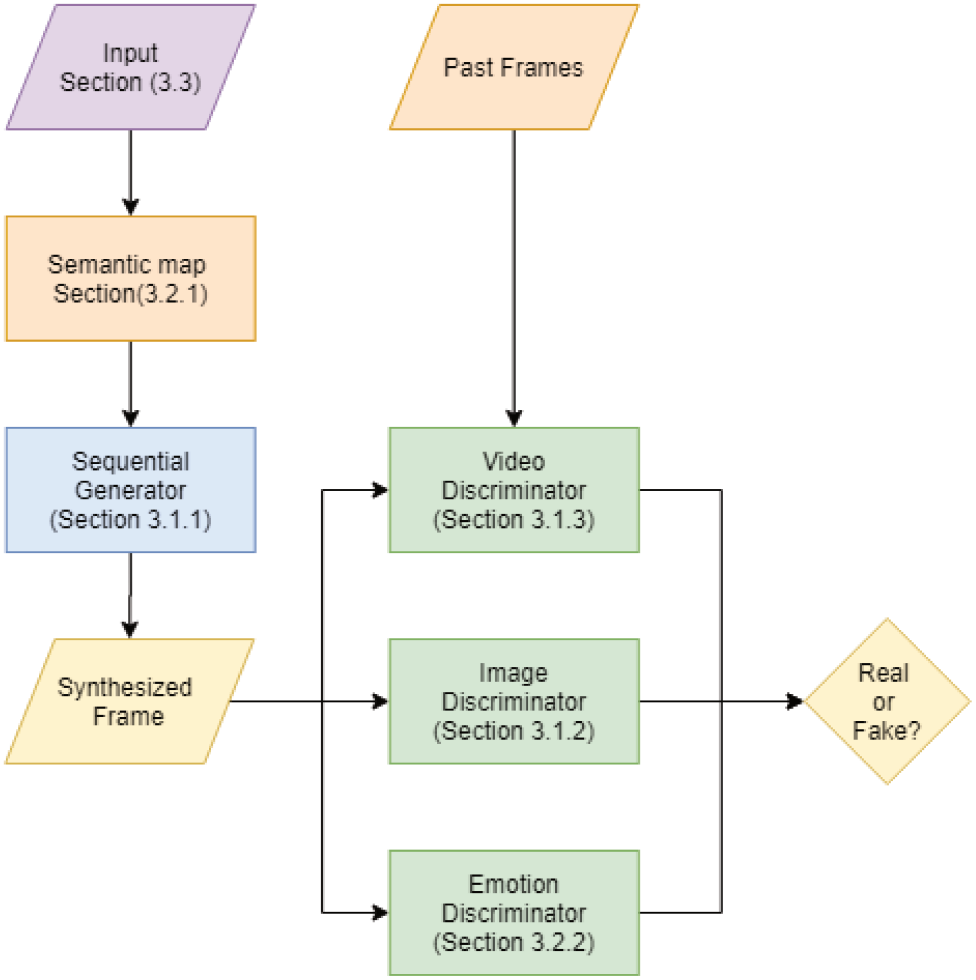


Figure 3.1 – Overview of our approach with the respective sections. Our approach uses semantic segmentation maps with information for both the target emotion and facial keypoints as input to a sequential generator. Additionally, this sequential generator also considers the past synthesized frames to generate new images. The frames synthesized are analyzed by three distinct discriminators the evaluate the temporal consistency (video discriminator), image quality (image discriminator), and the target emotion (emotion discriminator).

where:

- $P(\hat{y}_1^K|x_1^K)$ is the conditional distribution of the **synthesized data** given the input segmentation maps;
- $P(y_1^K|x_1^K)$ is the conditional distribution of the **real data** given the input segmentation maps.

The *vid2vid* approach to the synthesis task is based on the assumption that video signals contain redundant information in consecutive frames, and that the next frame is a function of its segmentation map given the previous frames. Hence, the generator used is considered a sequential generator. In particular, given the significant redundancy

on video signals, Wang et al. (2018) proposes the use of a window of 2 past frames for optimal training stability and resource consumption, as presented on Equation 3.2:

$$P(\hat{y}_1^K|x_1^K) = \prod_{k=1}^{K} P(y_k|y_{k-2}^{k-1}, x_{k-2}^k), \tag{3.2}$$

where:

- $P(\hat{y}_1^K|x_1^K)$ is the conditional distribution of the **synthesized data** given the input maps;
- $\prod_{k=1}^{K} P(y_k|y_{k-2}^{k-1}, x_{k-2}^k)$ is the product of the conditional distribution of the **real data** given the input maps for the window of 2 frames;

To compose the sequential generator, the authors propose using an optical flow calculation between frames to estimate the next image. Wang et al. (2018) base this on the assumption that the next frame may be obtained by warping the current frame with the optical flow between the past consecutive frames. The estimated optical flow is represented by Equation 3.3:

$$\hat{w}_{k-1} = W(\hat{y}_{k-2}^{k-1}, x_{k-2}^k), \tag{3.3}$$

where:

- $\hat{w}_{k-1}$ is the estimated optical flow;
- W is the optical flow network;
- $\hat{y}_{k-2}^{k-1}$ are the past synthesized images;
- x_{k-2}^k are the past segmentation maps.

This composition is achieved through the process mapped on Equation 3.4 [1]:

$$G(\hat{y}_{k-2}^{k-1}, x_{k-2}^k) = (1 - \hat{m}_k) \odot \hat{w}_{k-1}(\hat{y}_{k-1}) + \hat{m}_k \odot \hat{h}_k, \tag{3.4}$$

where:

- G is the function associated with our complete generator;
- $\odot$ is the element-wise product operator;

[1] The piece of code responsible for implementing Equation 3.4 is available at <https://github.com/fireis/emot_vid2vid/blob/master/models/networks.py#L229>

- $\hat{w}_{k-1}$ is the estimated optical flow;
- $\hat{y}_{k-2}^{k-1}$ are the past synthesized images;
- x_{k-2} are the past the semantic segmentation maps;
- $\hat{h}_k = H(\hat{y}_{k-2}^{k-1}, x_{k-2}^{k})$ is the intermediate image synthesized directly by the generator, noted as H;
- $\hat{m}_k$ is an occlusion mask used to handle background information, being M the mask prediction network.

We represent this process graphically in Figure 3.2. The optical flow prediction network uses the image generated by the conditional generator and the past frames to estimate a corresponding flow map. The next step in the process is to warp the flow map into the past synthesized frame, generating a warped frame. The last step in the process is to combine the image synthesized by the conditional generator and the warped frame, resulting in a time consistent frame.

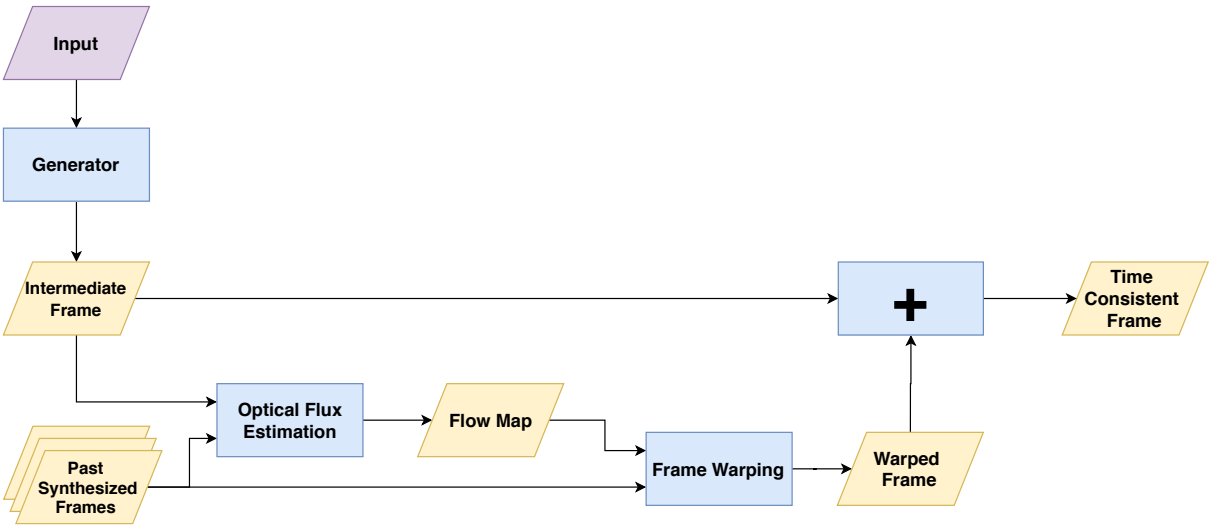


Figure 3.2 – The structure of the *vid2vid* sequential generator. The conditional generator uses semantic segmentation maps as input to generate an intermediate frame. This frame and the past synthesized frames are fed into a flow network to generate a flow map, which is warped to the previous frame, generating a warped frame. Finally, the warped frame is combined with the intermediate frame produced by the generator and result in a time consistent frame.

3.1.2 Image Discriminator

To evaluate the image quality of the results produced by the sequential generator, Wang et al. (2018) uses a multi-scale *PatchGAN* architecture for the discriminator. The *PatchGAN* discriminator aims to analyze the images on the scale of patches, verifying if

patches of a given size are real or fake, and its structure is shown in Figure 3.3 (ISOLA et al., 2017). The multi-scale aspect of the discriminator used on *vid2vid* refers to a combination of different *PatchGANs* changing the number of intermediate channels analyzed on the network, as shown on Figure 3.4, where an image is analyzed.

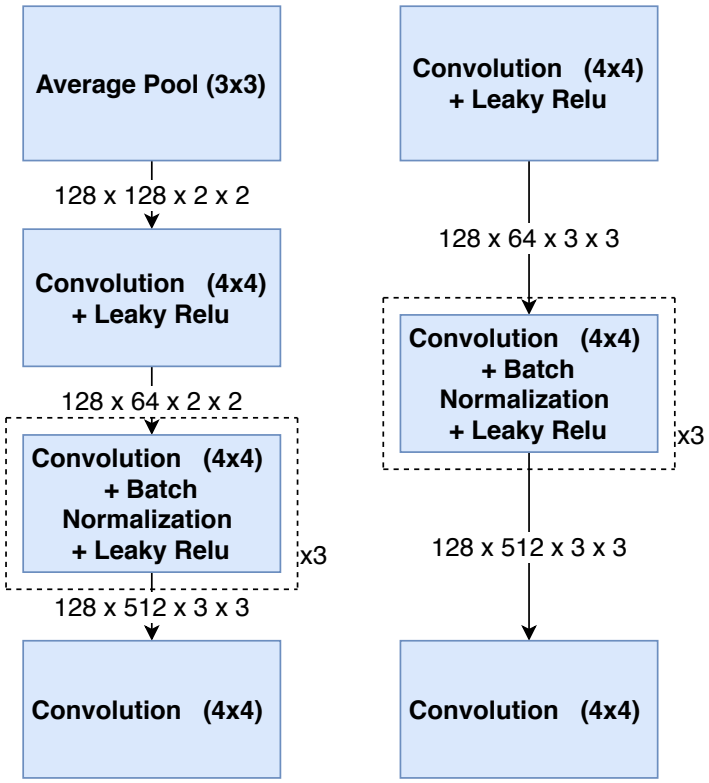


Figure 3.3 – Representation of *PatchGAN* discriminator structure proposed by Isola et al. (2017). This discriminator analyzes the images at the scale of patches, and classifies if each patch is real or fake. After assessing the whole image through patches, an average is calculated to obtain the final output of the discriminator.

3.1.3 Video Discriminator

To evaluate the video quality, *vid2vid* uses a multi-scale video discriminator. This discriminator also relies on a multi-scale PatchGAN network but validates if the generated frames are similar to real ones considering the optical flow. Additionally, to assure both short and long term temporal consistency, the video discriminator is temporally multi-scale. Wang et al. (2018) implement this multi-scale aspect by using a subsampling technique of skipping a given number of frames. In the finest scale, the discriminator analyzes every frame. On the other scales, the discriminator skips some of the previous frames, as presented in Figure 3.5. The authors propose up to three temporal scales.

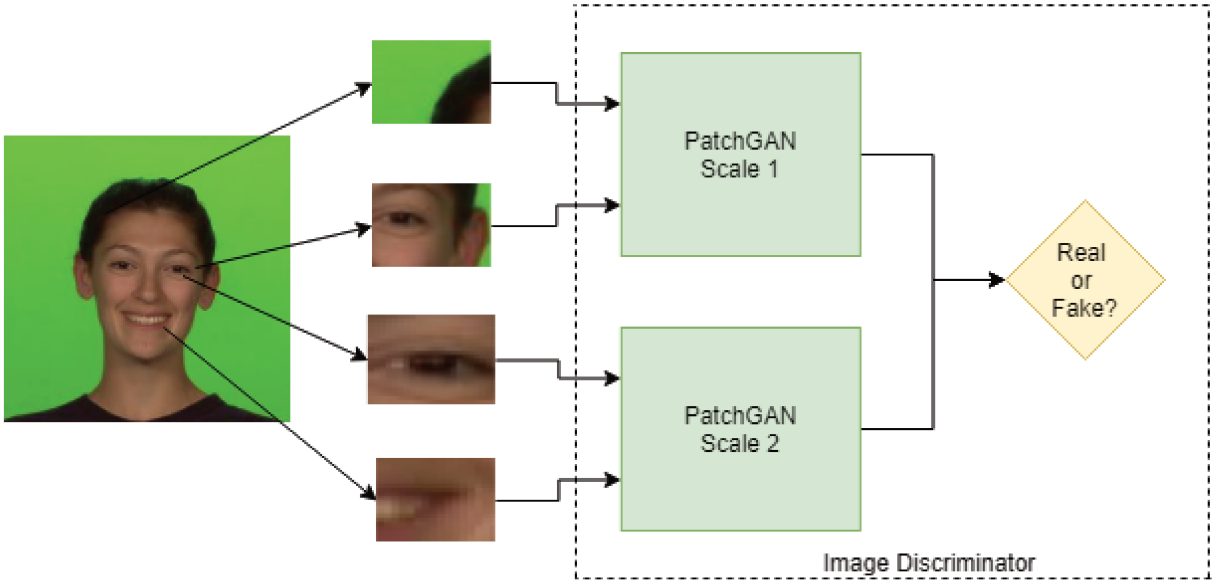


Figure 3.4 – Representation of the multiscale PatchGAN discriminator used in *vid2vid*. The PatchGAN discriminators evaluate the image divided on patches, evaluating if each patch is real or fake. The multiscale aspect of this discriminator is associated to the different patch sizes each network considers. We represent PatchGAN Scale 1 with a bigger patch size than PatchGAN Scale 2 to demonstrate this multiscale aspect.

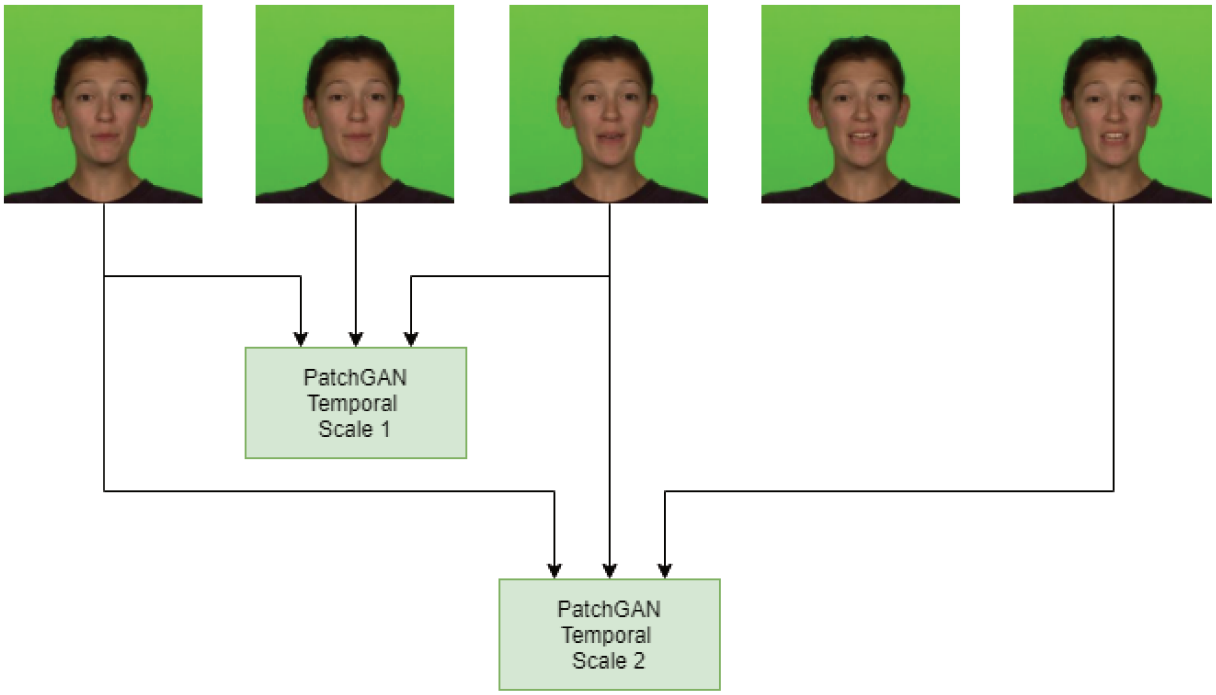


Figure 3.5 – Representation of the temporal multi-scale PatchGAN video discriminator used on *vid2vid*. Given a sequence of frames, the discriminator with the finest scale analyzes every frame, while the other scale always skips 1 frame. This approach allows the discriminator to better evaluate the video dynamics, penalizing sudden, unrealistic, changes on the sequence of frames.

3.1.4 Learning Objective

In Section 2.1 we presented the learning objective for a generic conditional GAN through Equation 2.6. To accommodate for the multiple discriminators and the sequential generator, we consider the adversarial loss indicated by $\mathcal{L}_{GAN}$ in Equation 3.5:

$$\mathcal{L}_{GAN} = \min_{G}(\max_{D_I}\mathcal{L}_I(G, D_I) + \max_{D_V}\mathcal{L}_V(G, D_V)) + \lambda_W\mathcal{L}_W(G), \tag{3.5}$$

where:

- $\mathcal{L}_{GAN}$ is the general adversarial loss of the network;
- $\mathcal{L}_W$ is the flow estimation loss;
- $\mathcal{L}_V$ is the video loss, associated with the video discriminator;
- $\mathcal{L}_I$ is the image loss, associated with the image discriminator;
- G is the complete generator;
- D_I is the image discriminator;
- D_V is the video discriminator;
- λ_W is a weight for the flow term.

The individual image loss may be represented by using the *Binary Cross Entropy Loss* from Equation 2.1, as shown in Equation 3.6:

$$\mathcal{L}_I = \mathbb{E}_{(y_1^K, x_1^K)}[\log D_I(y_i, x_i)] + \mathbb{E}_{(\hat{y}_1^K, x_1^K)}[\log(1 - D_I(\hat{y}_i, x_i)], \tag{3.6}$$

where:

- $\mathcal{L}_I$ is the image loss;
- D_I is the image discriminator;
- $\mathbb{E}_{(y_i^K, x_i^K)}$ is the expected value over the real data y_i^K instances given the corresponding segmentation maps x_i^K;
- $\mathbb{E}_{(\hat{y}_i^K, x_i^K)}$ is the expected value over the synthesized data, $\hat{y}_i^K$, given the segmentation maps x_i^K.

The video component of the loss obtained considering a window of N consecutive frames may also be obtained from Equation 2.1, as represented in Equation 3.7:

$$\mathcal{L}_V = \mathbb{E}_{(w_1^{K-1}, y_1^K, x_1^K)}[\log D_V(y_{i-N}^{i-1}, y_{i-N}^{i-2})] + \mathbb{E}_{(w_1^{K-1}, \hat{y}_1^K, x_1^K)}[\log(1 - D_V(\hat{y}_{i-N}^{i-1}, w_{i-N}^{i-2})], \quad (3.7)$$

where:

- $\mathcal{L}_V$ is the video loss;
- D_V is the video discriminator;
- w_1^{K-1} is the optical flow calculated for the past frames;
- $\mathbb{E}_{(w_1^{K-1}, y_1^K, x_1^K)}$ is the expected value over the real data, y_1^K, given the optical flow calculated w_1^{K-1} considering the past real frames, and the segmentation maps x_1^K;
- $\mathbb{E}_{(w_1^{K-1}, \hat{y}_1^K, x_1^K)}$ is the expected value over the synthesized data, $\hat{y}_1^K$, given the optical flow calculated considering the past synthesized frames, and the segmentation maps x_1^K.

Finally, the flow loss is presented in Equation 3.8, as defined by Wang et al. (2018):

$$\mathcal{L}_W = \frac{1}{K-1} \sum_{k=1}^{K-1} (\|\hat{w}_k - w_k\|_1 + \|\hat{w}_k(y_k) - y_{k+1}\|_1), \quad (3.8)$$

where:

- $\|\|_1$ represents the L1 norm;
- w_k is the ground truth flow calculated using the original frames from y_k to y_{k+1};
- $\hat{w}_k$ is the flow calculated using the synthesized frames from $\hat{y}_k$ to $\hat{y}_{k+1}$;
- $\|\hat{w}_k - w_k\|_1$ represents the error between the ground truth and the estimated flow;
- $\|\hat{w}_k(y_k) - y_{k+1}\|_1)$ represents the warping loss obtained by the flow warping the previous and the next frames.

The flow loss accounts for the error between the original and the synthesized flow, and the error associated with the flow warping with the previous frame.

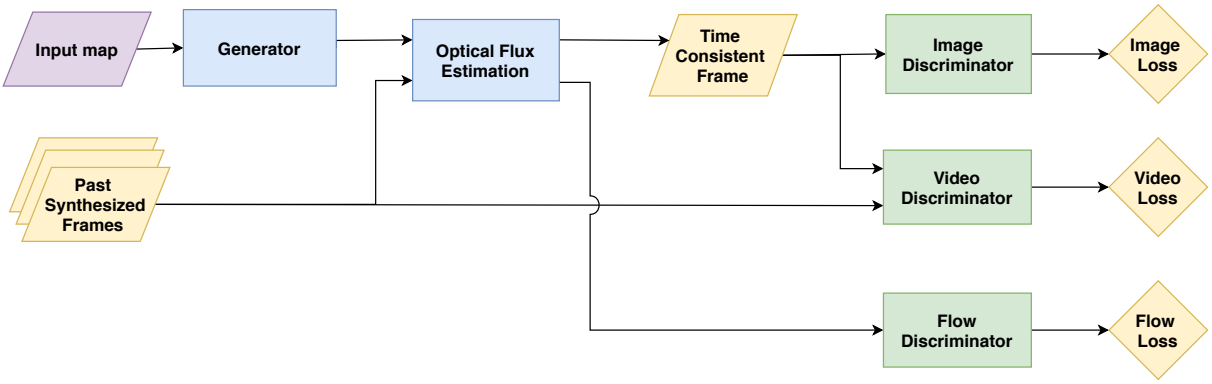


Figure 3.6 – Representation of the composition of the losses. The image loss is calculated using the synthesized temporal consistent frame. The video loss is calculated evaluating both the past frames and the temporal consistent frame. We calculate the flow loss by evaluating the flow maps from past and current frames.

3.1.5 Network Training

The *vid2vid* network can produce high-resolution results, of up to 2048 x 1024 using the coarse-to-fine strategy of the generator. This high resolution requires an impressive computational power, taking ten days to train the system using a *Nvidia DGX1* machine. Due to our computational power limitations, we use the network on lower resolutions, which do not require the use of multiple generators. We have used a Linux server with 1 Nvidia V100 GPU, eight processor cores, and 32 Gb of RAM to run our tests. The network consumed around 13Gb of video memory in the training process, considering a training dataset with 166 seconds of video at 256 per 256 pixels. In this configuration, each epoch took an average of 20 minutes to complete.

3.2 Our Network

In this section, we present the key contributions that compose our approach. We start by presenting the semantic segmentation proposed. Then we present the strategy for the emotion discriminator and the learning objective.

3.2.1 Semantic Segmentation Map

To generate expressive speech, we need to provide information relative to the target speech to our network. The input to the original *vid2vid* network is an image with a drawing related to the facial keypoints with a black background and white lines. The generator processes this entire image. One option to consider the target emotion information would be to use a second generator to generate a facial expression with the desired emotion and warp this with the main sequential generator's result. Although valid,

this approach increases the computational cost of the network. To avoid an increase in the computational cost and better optimize the structure already available on the sequential generator, we propose embedding the target emotion information on the color layer of the generator input. We do this by associating the background color and face line color with the target emotion. We use only greyscale colors to avoid increasing the complexity of the generator. Figure 3.7 shows the inputs for different emotions using our approach.

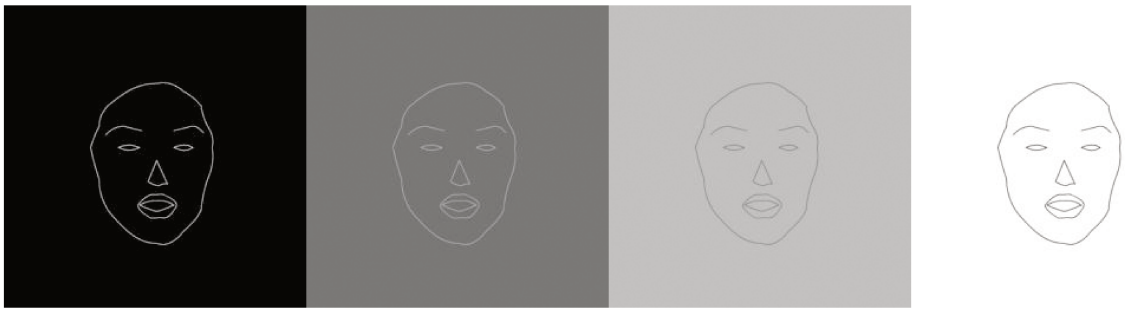


Figure 3.7 – Semantic segmentation maps for different emotions used as input to our network. The target emotion defines both the background color and the contour line color. The emotions associated with the colors are, from left to right, *Happy-for*, *Admiration*, *Fear*, and *Anger*.

With the emotion information, we describe the new objective for the generator as in Equation 3.9. As we add the emotion layer, we need to add another term, e, to account for the emotion target, resulting in the following equation:

$$P(\hat{y}_1^K|(x_1^K, e_1^K)) = P(y_1^K|(x_1^K, e_1^K)), \tag{3.9}$$

where:

- $P(\hat{y}_1^K|(x_1^K, e_1^K))$ is the conditional distribution of the **synthesized data** given the input maps (x_1^K) and the target emotion (e_1^K);
- $P(y_1^K|(x_1^K, e_1^K))$ is the conditional distribution of the **real data** given the input maps (x_1^K), and the target emotion (e_1^K).

3.2.2 Emotion Discriminator

By providing input regarding a target emotion, our system may associate said inputs to each emotion's facial expression. However, we do not have a structure to monitor and penalize our network if the synthesized images do not contain the expressions related to these emotions. To guide the improvement of the network during the training stage, we propose a dedicated emotion discriminator.

We propose using a *PatchGAN* discriminator dedicated exclusively to verifying if the emotion associated with the facial expressions on the synthesized videos is the same

as defined by the target. The objective of this discriminator is to assess if the perceived emotion on the resultant image follows the target emotion. The objective function, based on the *Binary Cross-Entropy Loss*, Equation 2.1, for this discriminator, awards images with the correct emotion and penalizes ones with the incorrect one, as described on Equation 3.10, where e represents the target emotion:

$$\mathcal{L}_E = \log(D_E(G(x|e))) + \log(1 - D_E(G(x|e))), \tag{3.10}$$

where:

- $\mathcal{L}_E$ is the emotion loss;
- D_E is the emotion discriminator;
- $G(x|e)$ is the result produced by the generator using the input segmentation map and the target emotion.

This additional discriminator is intended only for emotion control. The loss of this discriminator is accounted in conjunction with the other discriminator losses described in Section 3.1. We have used the same discriminator structure as the image discriminator, shown in Figure 3.3. The final loss for our approach, considering this new element, is presented in Section 3.2.3.

3.2.3 Learning Objective

The learning objective of our approach builds on the one from *vid2vid*, previously presented on Equation 3.5. The new learning objective also accounts for the emotion discriminator term, as shown in Equation 3.11:

$$\mathcal{L}_{GAN} = \min_G(\max_{D_I}\mathcal{L}_I(G, D_I) + \max_{D_E}\mathcal{L}_E(G, D_E) + \max_{D_V}\mathcal{L}_V(G, D_V)) + \lambda_W\mathcal{L}_W(G), \tag{3.11}$$

Where:

- $\mathcal{L}_{GAN}$ is the general adversarial loss of the network;
- $\mathcal{L}_W$ is the flow estimation loss;
- $\mathcal{L}_V$ is the video loss, associated with the video discriminator;
- $\mathcal{L}_I$ is the image loss, associated with the image discriminator;
- $\mathcal{L}_E$ is the emotion loss, associated with the emotion discriminator;
- G is the complete generator;

- D_I is the image discriminator;
- D_V is the video discriminator;
- D_E is the emotion discriminator;
- λ_W is a weight for the flow term.

3.2.4 Network Training

Our approach's network training follows the training from the original *vid2vid* network, as presented in Section 3.1.5. The difference between ours and *vid2vid* is the addition of the emotion label map and the emotion discriminator. Even though we have added the discriminator, both the time to complete each epoch and the memory consumed did not change compared to the original *vid2vid* network.

3.3 Image preprocessing

Our system uses a sequence of facial keypoints to generate a drawing of the face to be synthesized. In this section, we present our approach to obtain the facial keypoints, and to transform these coordinates into images.

3.3.1 Facial Keypoint Identification

Proper facial keypoint detection is paramount to achieving a good result in our system, as we use these keypoints to generate the input to our system. The facial keypoint identification problem is well known, with standard software packages providing excellent results with little computational power.

In the development stage, we have tested different approaches to identifying facial keypoints. The two approaches that were more extensively tested were *DLIB* (KING, 2009), and the *face align* method (BULAT; TZIMIROPOULOS, 2017). Both approaches generate a sequence of 68 facial keypoints, as shown in Figure 3.8.

The *DLIB* approach provided good results when the head of the actress did not present any rotation. In contrast, when the actress performed small, natural, facial rotations while speaking *DLIB* failed to detect the keypoints on some frames.

The *face align* approach achieves better results to identify the facial keypoints even with slight facial rotations. Due to this improved performance, we use the *face align* approach to obtain the facial keypoints used in our experiments.

3.3.2 Semantic Segmentation Map Generation

To generate the semantic segmentation map images used as input to our system, we use the 68 facial keypoints detected as described in Section 3.3. As these keypoints describe only the region from the jaw to the eyebrows, Wang et al. (2018) propose estimating the forehead region using the Equation 3.12 to achieve better mapping performance:

$$k_n = \hat{c_y} + round((\hat{c_y} - c_{y_n}) * 2/3), \tag{3.12}$$

where:

- k_n represents a forehead point, with n ranging from 2 to 13 (see Figure 3.8);
- $\hat{c_y}$ is the average vertical coordinate of the face, estimated using the average vertical coordinate from keypoints 1 and 14;
- c_{y_n} is the vertical coordinate for each of the facial keypoints from 2 to 13.

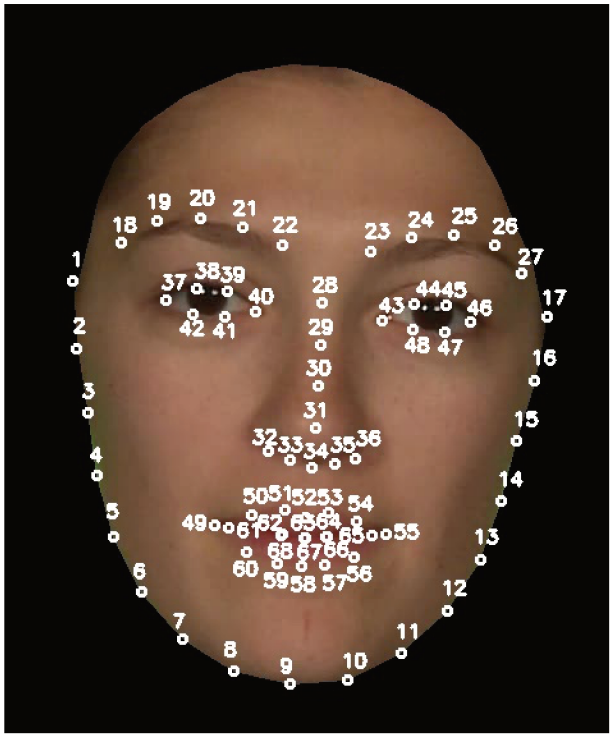
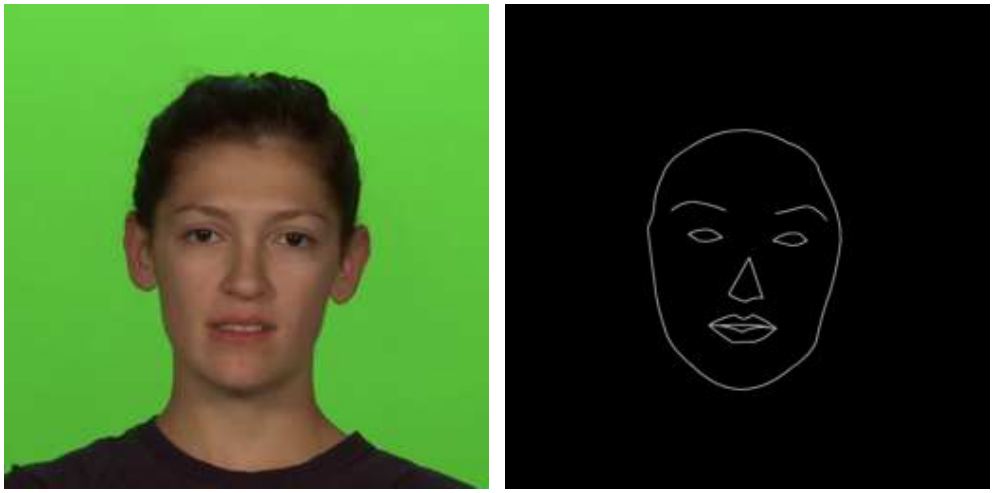


Figure 3.8 – Facial keypoints used to generate the input map to our system. We use the *face align* approach to detect the facial keypoints on our original frames as this approach achieves good and stable results even when the actors rotate their faces during the speech.

With the forehead points considered, Wang et al. (2018) propose the interpolation of near points generation lines that represent the shapes of the face, as presented in Figure 3.9.



(a) Ground truth image. (b) Semantic segmentation maps.

Figure 3.9 – Original frame from the dataset proposed by Costa (2015) and the corresponding semantic segmentation map used as input to our network.

We obtain the facial keypoints and generate the semantic segmentation map for every frame in our dataset. As presented in Section 3.2.1, the semantic segmentation map background color, and the color of the facial shape lines are defined by the emotion associated with the original video.

3.4 Objective results

To evaluate our results in an objective manner, we adopted the *FID* score.The *InceptionV3* network used to calculate the *FID* of our results was trained on the *ImageNet* train set. To provide a reference to the *FID* score, Figure 3.10 shows results obtained by Heusel et al. (2017).

Although *FID* can provide a direction in the decision making process to improve the results, the metric alone cannot provide definitive answers on the perceived videorealism of the results. In Figure 3.10, we can see that a face with a highly distorted shape presented in the bottom left corner has a better score (around 150) than a face with the correct shape but noisy (around 250), on the upper right corner. This indicates that the evaluation performed by the *FID* score may differ from the human perception, justifying the perceptual evaluation performed to assess the quality of our work, presented in Chapter 4.

The results were generated after training our model for 160 epochs, with an output format of 256 by 256 pixels. The results are presented in Table 3.1.

The results for the *FID* score shows good indication that our approach outperforms *vid2vid*, since the smaller the score, the better. This result indicates that our approach tends to synthesize images with a distribution more similar to the distribution

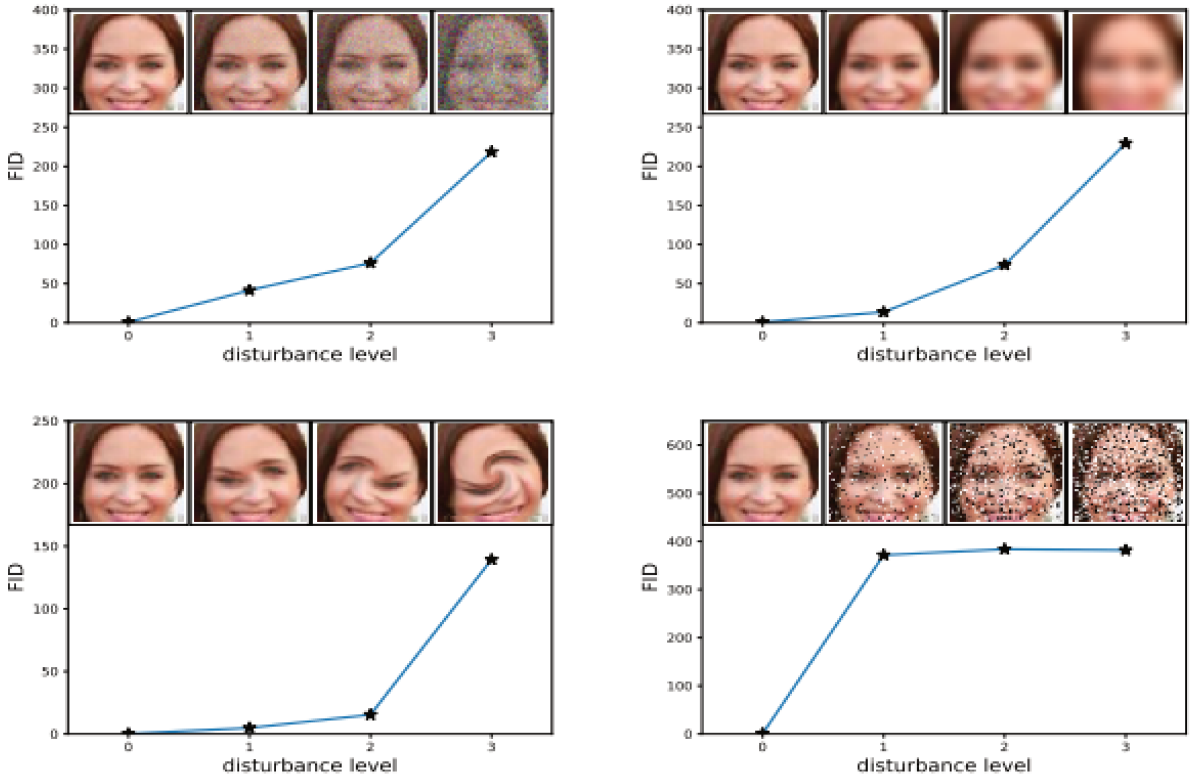


Figure 3.10 – *FID* scores for different perturbations applied to the original image. We verify that the *FID* score is lower to instances when the texture of the image is degraded, such as the one in the bottom right corner. In contrast, in the example of in the bottom left side where the texture is almost intact but the shape is distorted the *FID* score is not as affected. Source: image extracted from (HEUSEL et al., 2017).

Approach	FID
Vid2vid	38
Our Approach	32

Table 3.1 – Objective results for the set synthesized using our complete network and the original *vid2vid* network. The set used in this analysis is the same set from which the samples were taken for the user group evaluation. We show that our approach outperforms *vid2vid* in every metric considered in this work.

(a) Original frame. (b) Synthesized frame using *vid2vid*. (c) Synthesized frame using our approach.

Figure 3.11 – Comparison between the original frame and the synthesized frames using the *vid2vid* network, and our approach. The *vid2vid* network uses only the keypoints from the original frame to synthesize new images, while our approach uses both the keypoints and a target emotion. This results on our results having results less similar to the original frame than *vid2vid*, as we focus on always conveying the target emotion.

from the ground truth images than *vid2vid*. The difficulty to assert this conclusion is due to the reduced set analyzed to calculate the *FID* due to limitations of the amount of reference keypoints.

Figure 3.11 exemplifies differences from results using ours and the *vid2vid* approach. The main difference is that while *vid2vid* synthesizes the resulting image considering only the input keypoints, our approach also makes an effort to maintain the resulting face with the target emotion, even if the actress on that specific frame was not conveying this emotion.

We consider our objective results positive compared to the *vid2vid* network because we achieve a better score on *FID*, which is a good measure of similarity to real images.

We obtained the results considered in this section by training our system using a Linux server with 1 Nvidia V100 GPU, eight processor cores, and 32 Gb of RAM. The training stage consumed approximately 48 hours to train until epoch 160.

3.4.1 Ablation Study

Ablation studies are vastly used in the development of neural networks to determine each component's influence over the final result. These studies typically consist of removing elements of a neural network and evaluating the results without each element. This study may be instrumental because multiple structures are tested simultaneously in the development of neural networks, which makes the mapping of the influence of each component a tough task.

An ablation study was performed to assess the influence of the different elements

that we have introduced to our network. To obtain these results, we have trained one model for each approach for 140 epochs. We have set the model's output as 128 per 128 pixels to reduce the computational power required to train the models.

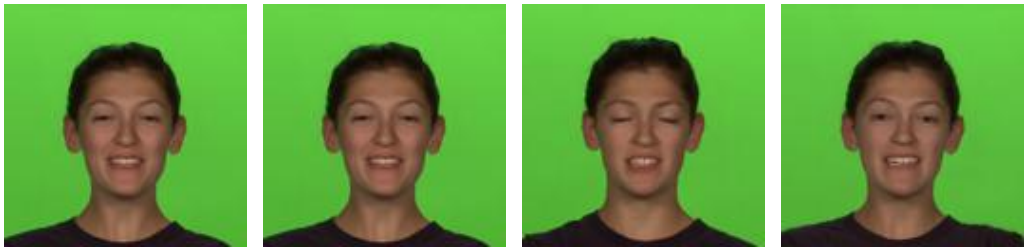
We extracted the inputs used in this test from videos where the actress performed the sentences with the *resentment* emotion, which presents a low level of activation and can be considered close to neutral as per empirical results presented on a previous study by Costa (2015). We want to evaluate if our system can generate results with different target emotions even without the emotion information in the segmentation maps, and the emotion discriminator.

We present the results in Table 3.2, where we calculated the *FID* score considering the synthesized results and original videos from the *Resentment* emotion. We also present sample frames obtained using each model and the same keypoints extracted from the *Resentment* emotion in Figure 3.12.

Approach	FID
Full system	51
Without Emotion Discriminator	52
Without Segmentation Map	53
Vid2vid	54

Table 3.2 – Results of our ablation test. We have trained four models with different network configurations to assess the influence of each component that we propose on our work. The first approach is the original *vid2vid* network. The second is the complete system, with all of the proposed components. Another approach is with the segmentation map described in Section 3.2.1 but without the emotion discriminator. The last approach is a network with the emotion discriminator described in Section 3.2.2, but without the segmentation map. The results present the positive influence of both elements as the full system is the approach with the best performance, which is measured by the lowest *FID*.

Analyzing the *FID* for each approach, we verify that we obtain better results when using our full proposed approach by comparing the scores from the *vid2vid* approach and our full approach. We can conclude that our proposed network can produce results with a distribution more similar to the distribution from the ground truth images. These realistic results may be due to the increase of information provided to our model, increasing its ability to generate better results. Another possible explanation for this improved performance is that by controlling the target emotion, we avoid emotion inconsistencies on the emotion expressed that occur when no emotion information is given. We use this result as an indication that using our full proposed approach optimizes our result. To measure the final quality of our results we consider the perceptual studies presented in Chapter 4.



(a) Original *vid2vid* network, FID 54.2. (b) Proposed network without the emotion discriminator, FID 52.2. (c) Proposed network without the segmentation map, FID 52.8. (d) Proposed network, FID 50.6.

Figure 3.12 – Results for ablation test considering the same input keypoints for each approach, obtained from a video of the actress performing the sentences with the *Resentment* emotion. The target emotion was set to *Happy-for*. We can see that the approach 3.12d presents the most realistic result, with both eyes open and mouth close to a smile.

3.4.2 Evaluation Across Epochs

We evaluate the evolution of the *FID* score as we train our model for more epochs. This evaluation is important as a form to validate the effect that increasing the training stage has on the final results. We present the results in Table 3.3 and show samples for each epoch are in Figure 3.13.

(a) Epoch 20. (b) Epoch 40. (c) Epoch 60. (d) Epoch 80.

(e) Epoch 100. (f) Epoch 140. (g) Epoch 160.

Figure 3.13 – Samples from the results obtained training our system for a different amount of epochs. As we increase the amount of epochs, the results become more realistic, eliminating artifacts that should not be present, such as head deformations encountered from epoch 20 to 60.

The analysis of the *FID* for each epoch demonstrates that we have obtained the

Epoch	FID
Epoch 20	114
Epoch 40	46
Epoch 60	40
Epoch 80	33
Epoch 100	34
Epoch 140	33
Epoch 160	32

Table 3.3 – Results obtained considering a set generated using our proposed model trained for a different amount of epochs. This result demonstrates that the training process indeed allows our system to improve the final result.

best result on epoch 160, which is the chosen epoch to generate the final results used on our user group evaluation. During the development stage of the network, we have stopped the training at epoch 160. We chose this stopping point due to the computational cost associated with the training, as at this stage of the training, each epoch took more than 20 minutes to be processed.

3.5 Concluding Remarks

In this chapter we presented the *vid2vid* network in Section 3.1, which is the basis for our approach. In Section 3.2, we presented the network we proposed, as well as our key contributions for adding emotion control to the *vid2vid* network. Our approach can generate consistent results without increasing the complexity and computational cost of the system. In Section 3.3, we presented the preprocessing stage to generate our training dataset. Finally, in Section 3.4, we presented the tests we performed during the development of our network. The results presented in this chapter allowed us to make important design options and indicated that our system generates good results, especially when comparing to the original *vid2vid* network. In the following chapter, we present our user study evaluation to assess if the users can identify emotions associated with the facial expressions synthesized, and if they prefer results obtained using the *vid2vid* network or ours, properly evaluation the quality of our results given the limitations of the *FID* score for this purpose.

4 Videorealism Assessment

The evaluation process for visual synthesis projects is challenging, mostly because the final goal is subjective: to measure how videorealistic are the results. We use the videorealism definition as the extent to which animation can be discerned from a real video by the spectator (MATTHEYSES; VERHELST, 2015). Therefore, we consider our results videorealistic if they may be classified as real videos by our viewer population. Related research usually evaluates the resulting systems in one or more of three forms, with perceptual, subjective, and objective measures (MATTHEYSES; VERHELST, 2015).

Objective evaluations consider the synthesized video signal's measurable characteristics, such as the difference between the synthesized and the ground truth videos. Typically, these metrics rely on a mathematical calculation to compare a set of results with a collection of original data by considering the actual data or its probability distribution. Related works on GANs consider *FID* score as at least one of the objective metrics as this score evaluates if the synthesized data has a distribution similar to the original data, as discussed in Section 2.1.1 (WANG et al., 2018; WANG et al., 2019; ZAKHAROV et al., 2019). In other approaches, researchers consider similarity metrics such as *SSIM*, as these allow a direct comparison to the ground-truth data (QIAO et al., 2018; ZAKHAROV et al., 2019; KIM et al., 2018). We have used these metrics to design and evaluate our work during the development stage, as already presented in Section 3.4.

Another method of evaluating the results of visual facial synthesis is through perceptual measures. In studies that use these metrics, the researchers present speech fragments to a group of test participants and ask them to perform a series of tasks related to the comprehension of the information (MATTHEYSES; VERHELST, 2015). A notable example of a perceptual test is the evaluation of the intelligibility of the speech, where the users need to define what they have comprehended from the given stimuli (COSTA; MARTINO, 2013; EZZAT; GEIGER; POGGIO, 2002; DEY; MADDOCK; NICOLSON, 2010).

Subjective tests are similar to perceptual evaluations because they employ a group of test subjects to observe a series of visual speech fragments, differing that these subjects present their opinion about the evaluated pieces. Related work on visual speech synthesis typically uses subjective tests to evaluate the research. One substantial downside of this approach is that even small errors in the speech synthesis usually significantly degrade the results (MATTHEYSES; VERHELST, 2015). An example of this group of tests is to present the user a synthesized video sequence and ask to evaluate how natural the video feels on a *Likert* scale (FRIED et al., 2019; KIM et al., 2018; FILNTISIS et al.,

2017; RUHLAND; PRASAD; MCDONNELL, 2017).

Our focus in evaluating our results is to assess if they are videorealistic even while conveying facial expressions related to a set of emotions. Although this is a challenging task, we proposed subjective tests in Section 4.2, along with the test protocol. We evaluate our results by comparing them to ones obtained using the original *vid2vid* approach, which generates viderealistic results (WANG et al., 2018). Additionally, we assess emotion perception, focusing both on the actual emotion perception and valence perception.

In this chapter, we describe and discuss the results of a subjective study performed to evaluate the animations achieved by applying the synthesis methodology presented in Chapter 3. We start by presenting qualitative results in Section 4.1, showing different use cases. Sections 4.2 to 4.5 present all stages of the user study proposed to evaluate our results, from design to actual results. The final section of the chapter discusses the different results.

4.1 Qualitative Assessment

In this section, we present the qualitative results of our work. To better organize this section, we divide it into two subsections regarding the origin of the keypoints. Section 4.1.1 presents results obtained with keypoints from the actress in the training set, while Section 4.1.2 presents results with keypoints from another actor.

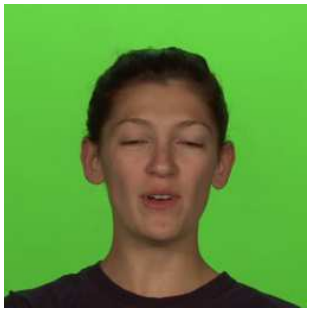
4.1.1 Keypoints from the Subject in the Training Set

To evaluate the quality of the synthesis using keypoints extracted from the actress present in the training set, we have used ground-truth videos of the emotion *Resentment*. In a previous study with the same dataset, users assigned mostly a neutral valence to this emotion (COSTA, 2015). These videos were not in the training set to impose a more significant challenge to our system. We present the results in Figure 4.1.

We note that the results are different for each emotion, showing that our system changes the output according to the emotion input. Additionally, we do not perceive any significant visual artifact in these samples.

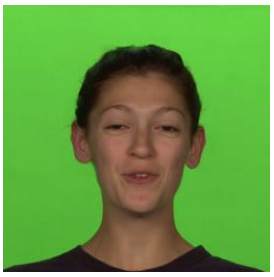
When analyzing more frames synthesized from keypoints extracted from videos of the actress in the training set, we note that sometimes the system generates frames with the subject blinking only one eye, as shown in Figure 4.2.

When we analyze the original image from which we obtained the keypoints, we note that the left eye of the actress was more closed than the other. Analyzing the outputs for every emotion, we note that our system has synthesized this difference in various intensities. With these differences, the only emotion where this is an actual quality

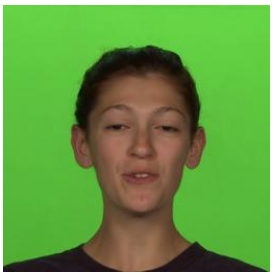


(a) Original sample.

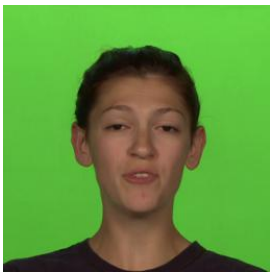
(b) Extracted keypoints.



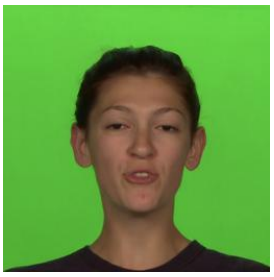
(c) Synthesis with *Happy-for* target.



(d) Synthesis with *Admiration* target.

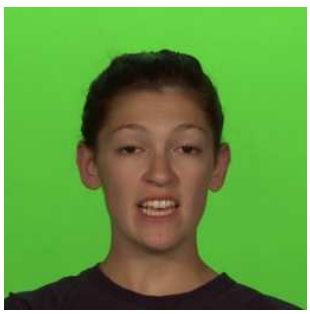


(e) Synthesis with *Fear* target.



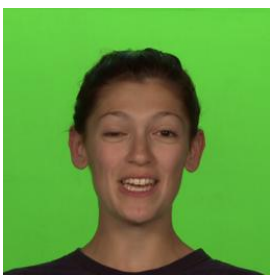
(f) Synthesis with *Anger* target.

Figure 4.1 – Original sample and corresponding synthesized frames. We generate one output for each emotion (c through f) by using as input the keypoints (b) extracted from the original frame (a)

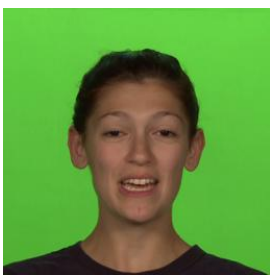


(a) Original sample.

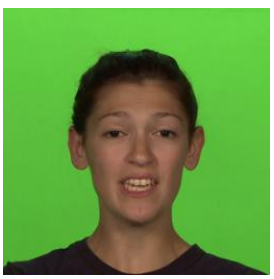
(b) Extracted keypoints.



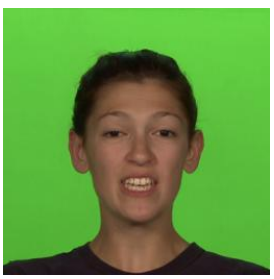
(c) Synthesis with *Happy-for* target.



(d) Synthesis with *Admiration* target.



(e) Synthesis with *Fear* target.



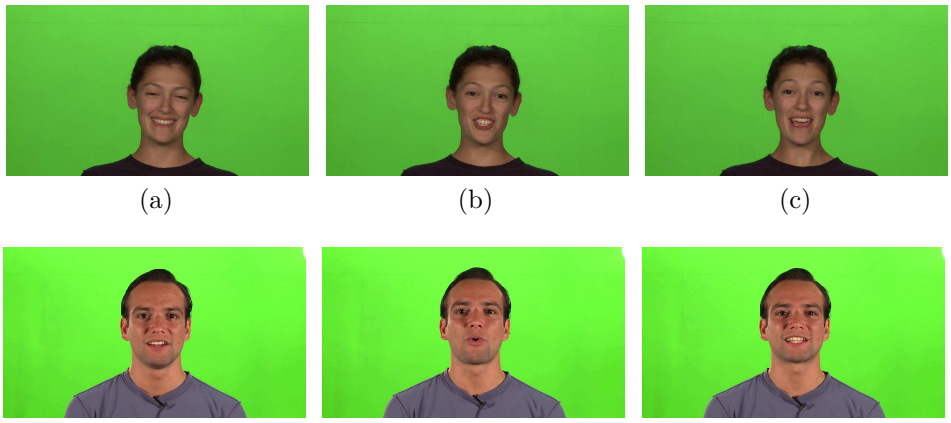
(f) Synthesis with *Anger* target.

Figure 4.2 – Original sample and corresponding synthesized frames. We generate one output for each emotion (c through f) by using as input the keypoints (b) extracted from the original frame (a). We note that our system synthesized frames with only one eye open for the *Happy-for* emotion.

issue is the *Happy-for*. This issue is likely associated with the system generating more intense expressions for this emotion and thus allowing such disturbance.

4.1.2 Keypoints from a New Subject

The speech is a natural process, and even with similar directions, each person has its way of expressing during the speech. This difference results in distinct head trajectories during the speech for each person. In Figure 4.3, we present frames recorded from original videos of each subject to present some differences in their head trajectories.



(a) (b) (c)

(d) (e) (f)

Figure 4.3 – Original samples from subjects both in the training set (a through c) and other subject (d through f). Both subjects were asked to express the same sentences and emotions. In the middle frames (b and e) we can note a difference in the head movement, with the actor making a more significant vertical head rotation than the actress.

The actress in the training set presents slight vertical head rotation, while the other actor moves its head in this direction in a more significant way. This difference results in significant discrepancies in the extracted keypoints, as presented in Figure 4.4.

We note that the shape of the head is significantly different when comparing the representation of the keypoints from both subjects. This variation is likely related to the vertical rotation of the head that the actor not in the training set (images d through f) performs when speaking. This difference imposes a challenge for our trained model as the samples in the training set did not cover head rotations with this intensity. The results for the synthesis of the keypoints in our analysis are presented in Figure 4.5.

Analyzing the synthesized images in Figure 4.5, we note that the results for the actress in the training set (a through c) are similar to original images, without visual artifacts that affect the quality, making them realistic. In contrast, the results for keypoints from another actor (d through f) are not as realistic. We note mainly two different artifacts: blurriness in the hair and forehead division, and visual artifact in the right portion of

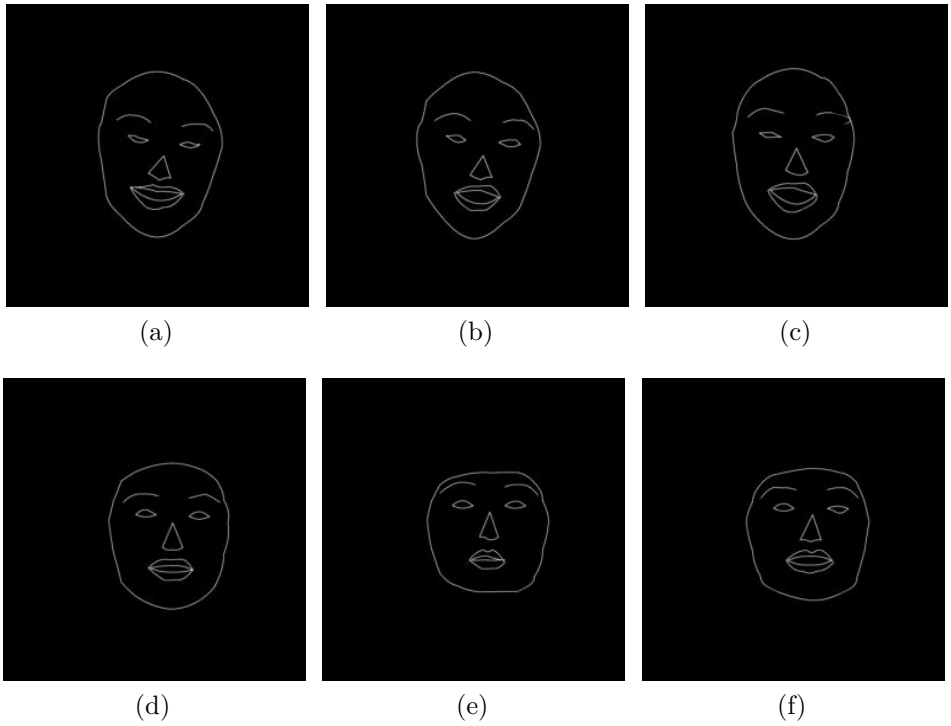


Figure 4.4 – Representation of the keypoints extracted from samples of subjects both in the training set (a through c) and a new subject (d through f). The keypoints were extracted using the same method for both subjects. The images from which the keypoints were extracted are presented in Figure 4.3.

the neck. We attribute both issues to the vertical movement of the head. When the actor rotates vertically rotates the head, the keypoints for its chin become higher than normal. This difference causes our model to consider the whole face shifted up and created the distortion on the neck region. Additionally, we attribute the forehead blurriness to the proximity between the eyebrow lines, and the forehead contour, creating difficulty for our system in this detection.

These visual artifacts represent the main challenge our system faces when generalizing input keypoints from other subjects. We could likely overcome this issue by either augmenting our training dataset, performing carefully defined operations in the original images to allow our model to learn new mappings. Another option to improve this would be to use in our training set a subject with more significant head movements, presenting more distinct movement options for our system during training.

4.2 Experiment Design

Our experiments were designed to evaluate if our system could generate results from which the users would be able to perceive the target emotion. To evaluate if our system was able to produce expressive results with the correct emotions, we had to define

Figure 4.5 – Original samples from subjects both in the training set (*a* through *c*) and other subject (*d* through *f*). Both subjects were asked to express the same sentences and emotions. In the middle frames (*b* and *e*) we can note a difference in the head movement, with the actor making a more significant vertical head rotation than the actress.

a set of target emotions. One common option of emotions choice is to adopt the six basic emotions proposed by Ekman (1993) (*Anger*, *Disgust*, *Fear*, *Happiness*, *Sadness* and *Surprise*). This approach may provide good results for the model in question, but are not an adequate representative for applications in the real world, as the most common emotions are typically more subtle (COSTA, 2015). Additionally, these six emotions are archetypal, which makes them easier for users to perceive correctly.

To impose a challenge resembling a real-world use case, we have adopted some emotions from the *OCC* model (ORTONY; CLORE; COLLINS, 1988). We have followed conclusions presented by Costa (2015) on the analysis of experimental results obtained with the dataset proposed by Costa (2015) to define the emotions used in our evaluation. These conclusions indicate that it is possible to cluster the 22 emotions in the *OCC* model into 5 different groups, *Strong Negative*, *Negative*, *Neutral*, *Positive*, and *Strong Positive*. We have chosen four emotions situated at the limits of each cluster: *Anger - Strong Negative*, *Fear - Negative*, *Admiration - Positive* and *Happy-for - Strong Positive*. The neutral emotion used as input was *Resentment - Neutral*. We consider this set an adequate representative of a real use case for the model. The ability to properly represent them is proof that our system can correctly express subtle emotions.

We performed the study with 42 participants, which had two tasks: (1) to determine the emotion shown in one video segment of a talking head, and (2) to choose the most realistic video between two choices. In every task, we presented only the video without the audio signal. These tasks were given on three different stimuli, as follows:

- **Neutral to Expressive Synthesis - Same Subject:** These stimuli consists of expressive facial animations synthesized from keypoints extracted from neutral speeches from the same actress from which we trained the face model, as presented in Figure 4.6. We extracted the keypoints from neutral speech videos (*Resentment* emotion) that were not present on the training set. Two animations were generated for each of four emotions (*Anger*, *Fear*, *Admiration* and *Happy-for*), totalling eight samples. In this test, we asked the participants to determine the emotion shown on each video;

- **Neutral to Expressive Synthesis - Different Subject:** These stimuli consist of expressive facial animations synthesized from keypoints extracted from neutral speeches from an actor whose face was completely new to the system. We extracted the keypoints from neutral speech videos (*Resentment* emotion) that were not present on the training set. We generated two animations for each of four emotions (*Anger*, *Fear*, *Admiration* and *Happy-for*), totalling eight samples, as presented in Figure 4.7. In this test, we asked the participants to determine the emotion shown on each video;

- **Synthesis Method Comparison:** These stimuli consist of pairs of expressive facial animations synthesized from keypoints extracted from neutral (*Resentment* emotion) speeches from the same actress from which we trained the face model. The difference between the videos in pairs is the synthesis method, as we synthesized one using our approach, and the other using the *vid2vid* network. We generated two animations pairs for each of four emotions (*Anger*, *Fear*, *Admiration* and *Happy-for*), totalling eight pairs of samples. In this test, we asked the participants to choose the most realistic video between the options.

Every participant in our study analyzed the three stimuli in a random sequence, as explained in Section 4.3.

4.3 Evaluation Protocol

The evaluation was conducted through a dedicated assessment application developed for this specific purpose[1]. This tool was necessary to provide all the flexibility and resources we needed for the planned study.

[1] The software for the evaluation study is available at <https://github.com/fireis/video-choices>

(a) Original frame from another actor with the *Resentment* emotion.

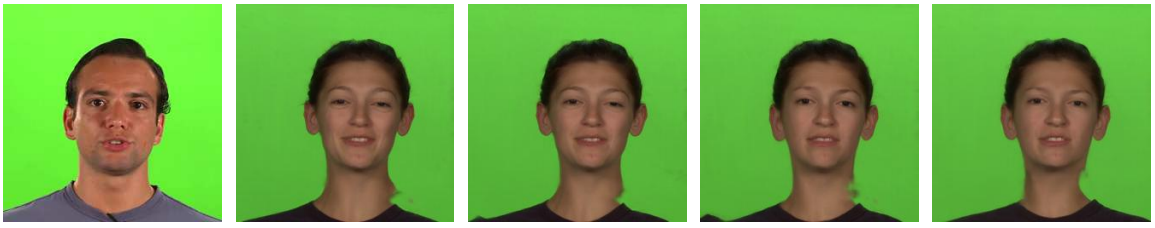
(b) Synthesized frame with *Happy-for* target.

(c) Synthesized frame with *Admiration* target.

(d) Synthesized frame with *Fear* target.

(e) Synthesized frame with *Anger* target.

Figure 4.6 – Example of original frame and corresponding synthesized results used on the neutral to expressive synthesis - same subject stimuli. In this test, we ask the users to chose the emotion they perceive in a video segment. We synthesized the video segments from keypoints obtained from an actor absent from the training set. We used two videos per emotion, totaling eight videos in this test.



(a) Original frame from another actor with *Resentment* emotion.

(b) Synthesized frame with *Happy-for* target.

(c) Synthesized frame with *Admiration* target.

(d) Synthesized frame with *Fear* target.

(e) Synthesized frame with *Anger* target.

Figure 4.7 – Example of original frame and corresponding synthesized results used on the neutral to expressive synthesis - different subject stimuli. In this test, we ask the users to chose the emotion they perceive in a video segment. We synthesized the video segments from keypoints obtained from an actor absent from the training set. We used two videos per emotion, totaling eight videos in this test.

The evaluation software allows two different evaluation mechanisms. The first process is when the user is presented with two videos and asked to choose one based on a given question, as presented in Figure 4.8. We used this process on the *neutral to expressive synthesis*, and *neutral to expressive synthesis* stimuli, for both same and different subjects. The second mechanism is to ask the user to choose between a set of alternatives based on a given question and a video, as shown in Figure 4.9. We have used this second mechanism to perform the *synthesis method comparison* test.

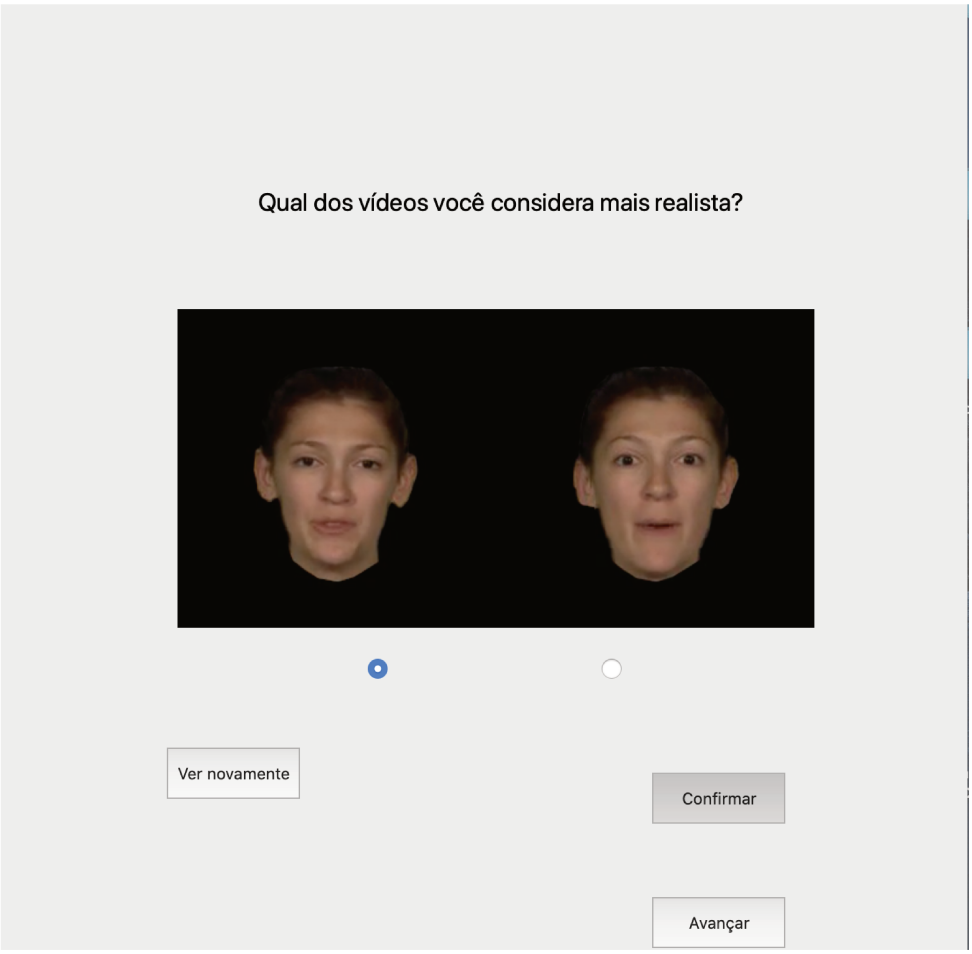


Figure 4.8 – Evaluation screen for stimuli comparing the most videorealistic of the two videos. In this screen, the user follows the flow described in Figure 4.10, with the main question being *"Which video do you consider most realistic?"* ("Qual dos vídeos você considera mais realista", in Portuguese) . The process starts with the videos playing simultaneously, and the user can restart the videos by clicking the *retry* button ("Ver Novamente", in Portuguese). After the video ends for the first time, the *confirm* button is made available ("Confirmar", in Portuguese). After the user confirms its choice, the *advance* button is enabled ("Avançar", in Portuguese). The user may watch the videos again or advance to the next stimuli. The analyzed samples are presented on each side of the screen an equal number of times and the default selected option is always the left one to avoid selection bias.

The evaluation process starts with questions regarding personal information,

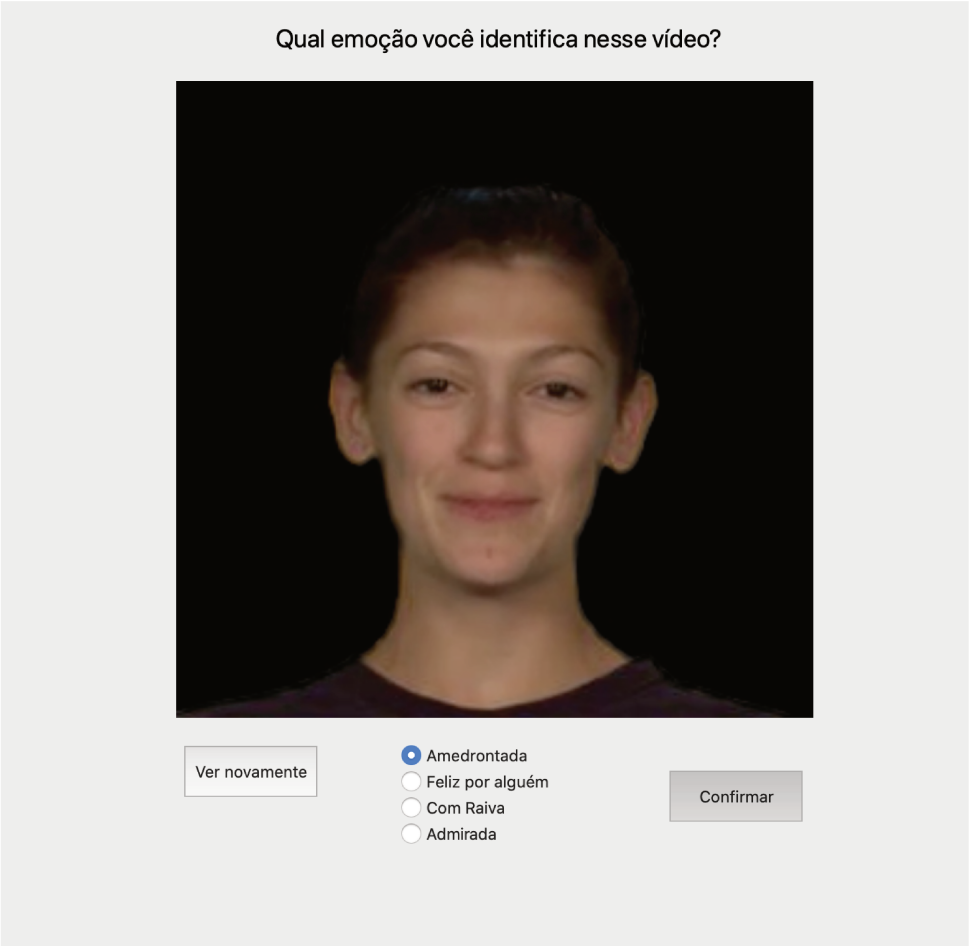


Figure 4.9 – Evaluation screen for emotion perception stimuli. The flow of this screen is described in Figure 4.10, and the main question is *"Which emotion do you identify in this video?"* ("Qual emoção você identifica nesse vídeo?", in Portuguese). After the screen loads, the user must watch the video until completion for at least one time. The user may restart the video from the beginning at any point. The four emotion options are randomly sorted, and the first option is always selected to avoid making the default selection improve the odds of choosing a single emotion. The four emotions presented are *Fear* ("Amedrontada", in Portuguese), *Happy-for* ("Feliz por alguém", in Portuguese), *Anger* ("Com Raiva", in Portuguese), and *Admiration* ("Admirada", in Portuguese). After the user chooses an option and clicks on the *confirm* button ("Confirmar", in Portuguese), the *advance* button is enabled ("Avançar", in Portuguese). If the user chooses to advance, the next evaluation screen is loaded.

and a box for the indication of compliance with the terms of the research[2].

After the initial screen, the software determines the next stimuli randomly. We use a random sequence of stimuli to avoid bias due to the order of the experiment. Using a random sequence of stimuli and different videos in each stimulus, we assure that each participant evaluates the options in a different sequence, reducing the influence of effects such as the subject being tired at the end of the test. Additionally, the order of the emotions presented in the screens of emotion perception is randomly defined. We have used this process to avoid selection bias due to the order of emotions presented. Finally, in the *synthesis comparison* test, we define randomly which video we present on the right and which is on the left of the screen to mitigate selection bias.

To evaluate the different stimuli, we use two types of screen, one for multiple videos and another for a single video. Both types of screens share the same flow presented in Figure 4.10. The process starts with the video playing, and the user can press a button to restart the video. After the video ends, a button to confirm the user choice is enabled, and the user may choose an option, and press *confirm* or replay the video as many times as desired. After the user confirms its option, a button to advance to the next step is enabled, and the user may either replay the video or advance to the next screen, ending the flow.

We conducted the test on a computer located in a closed room without external distractions.

4.4 Participants Profile

Forty-two subjects participated as volunteers in our user study. These users had no prior knowledge of our research or test purposes. None of the participants informed having visual impairment conditions, which could undermine their participation. We have considered data from every participant in our study. The age range of most participants was 26 to 35 years, as shown in Figure 4.11.

Our user study participants were mostly undergraduate, and graduate students of the University of Campinas, and the majority of them had already achieved higher education levels, as seen in Figure 4.12. We relate this to our user base being familiar with technology, eliminating distortions due to difficulties to complete the evaluation.

The users were closely distributed in the masculine and feminine gender, as shown in Figure 4.13.

[2] Consent term as per approval of the Ethics Committee of Unicamp, process *65289517.9.0000.5404*, available at <https://drive.google.com/file/d/10p9Ccs1FTF1TS3niE-boLQ-x3zU3iMrW/view?usp=sharing>

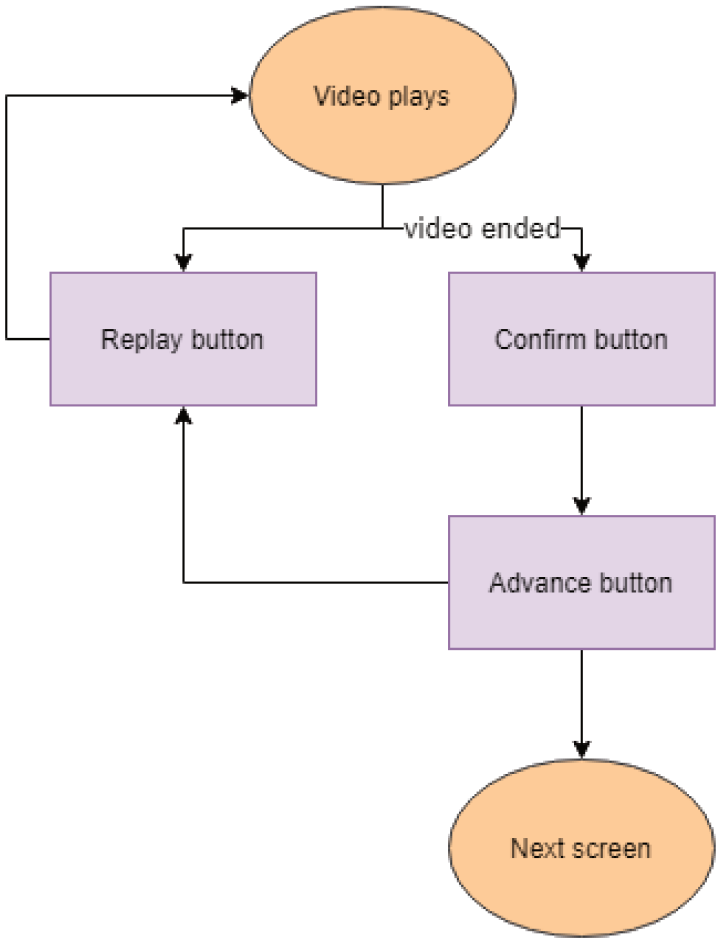


Figure 4.10 – Flow for the evaluation process. After the screen loads, the video starts playing. If the user presses the *replay* button, the video starts from the beginning. If the *retry* button is not pressed and the video ends, the *confirm* button is enabled. After the user presses the *confirm* button, the *advance* button is enabled. The user is still able to press the retry button and opt for another option. If the user presses the *advance* button, a new screen is loaded, and the same process is applied for the new evaluation instance.

4.5 Assessment Results

In this section, we present and discuss the results obtained for both the *Neutral to Expressive*, and the *Synthesis Method Comparison* evaluations.

4.5.1 Emotion Perception - Neutral to Expressive Synthesis

We evaluate the results of the emotion perception study by analyzing the answers the participants chose to each stimulus. In **neutral to expressive synthesis - same subject** stimuli, we aim to evaluate if our system can transform neutral inputs into expressive outputs. In the **neutral to expressive synthesis - different subject** stimuli, we want to evaluate if our system can generate expressive output and if our system can generalize the input face, as we use the facial keypoints from a new actor. As presented

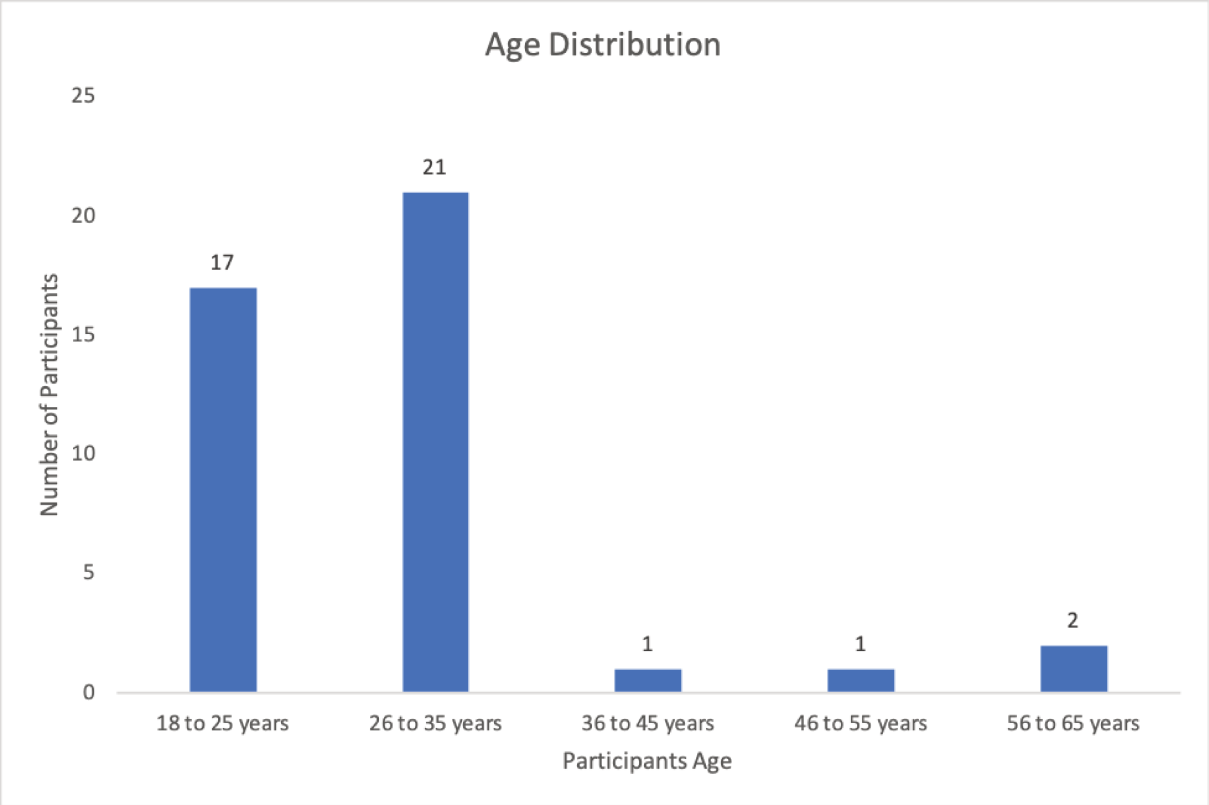


Figure 4.11 – Age distribution of the participants of the study. Most of the users are less than 35 years old, with the most significant range being 26 to 35 years.

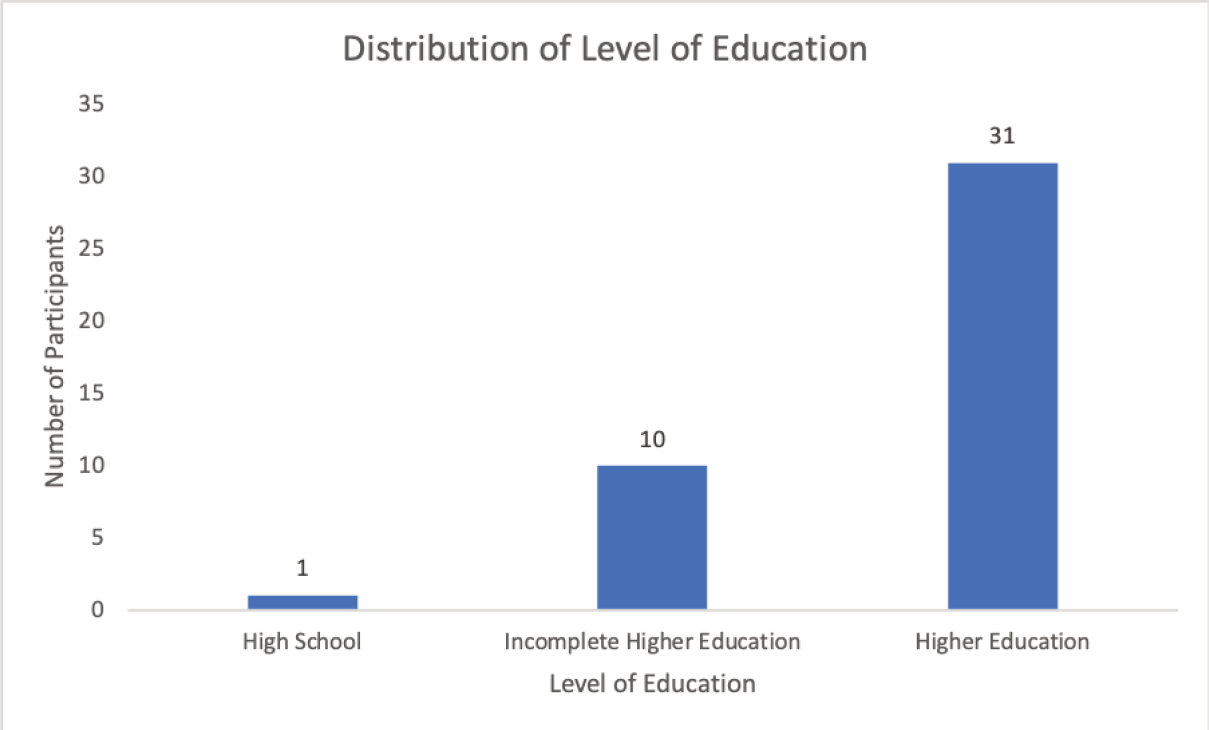


Figure 4.12 – Distribution of the levels of education of the users that participated in our study. The majority of the users had Higher Education, while there were approximately a quarter users that were either still on college or have left it and a single user with high school education.

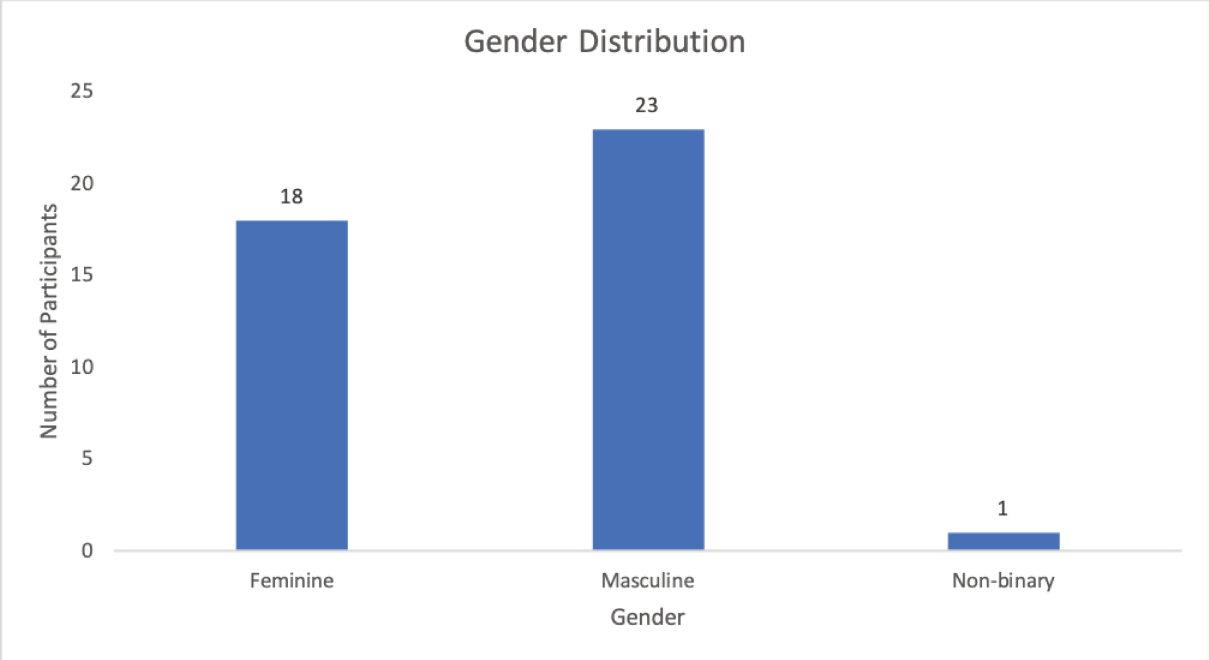


Figure 4.13 – The gender distribution for the user population. The masculine and feminine genders are well distributed, being 43% identified as feminine, and 55% as masculine.

in Section 4.5, using the keypoints from a different actor poses great difficulty to our system because the head movement and mouth articulations follow a different pattern than the ones presented by the original actress. We evaluated every result in the **Neutral to Expressive Synthesis** test using a *chi-squared* test, and confirmed that these results are statistically significant, with every p-value of $p < 10^{-2}$ or smaller.

Figure 4.14 shows the results for both **neutral to expressive synthesis** same and different subject stimuli considering videos generated with the *Happy-for* target. For this emotion, 81.7% of the participants chose the correct emotion on the **same subject** stimuli and 80.5% on the **different subject**. In contrast, 17.1% chose another emotion within the positive valence, resulting in a total of 98.8% perception as a positive valence.

The results for the stimuli considering video synthesized with the *Admiration* target are presented on Figure 4.15. In 35.4% of the instances of **neutral to expressive synthesis - same subject**, the participants perceived the target emotion, while on the different subject tests, this reduces to 30.5%. In contrast, 98.8% of the participants have chosen an emotion within the correct valence for the same subject test, which indicates that the system is capable of synthesizing video with the correct valence, but shows a deviation regarding the intensity. This difference is likely related to the emotion being more subtle than *Happy-for*. In **neutral to expressive synthesis - different subject**, 67.1% of the choices corresponded to emotion with a positive valence, which demonstrates that these samples caused more confusion on the participants.

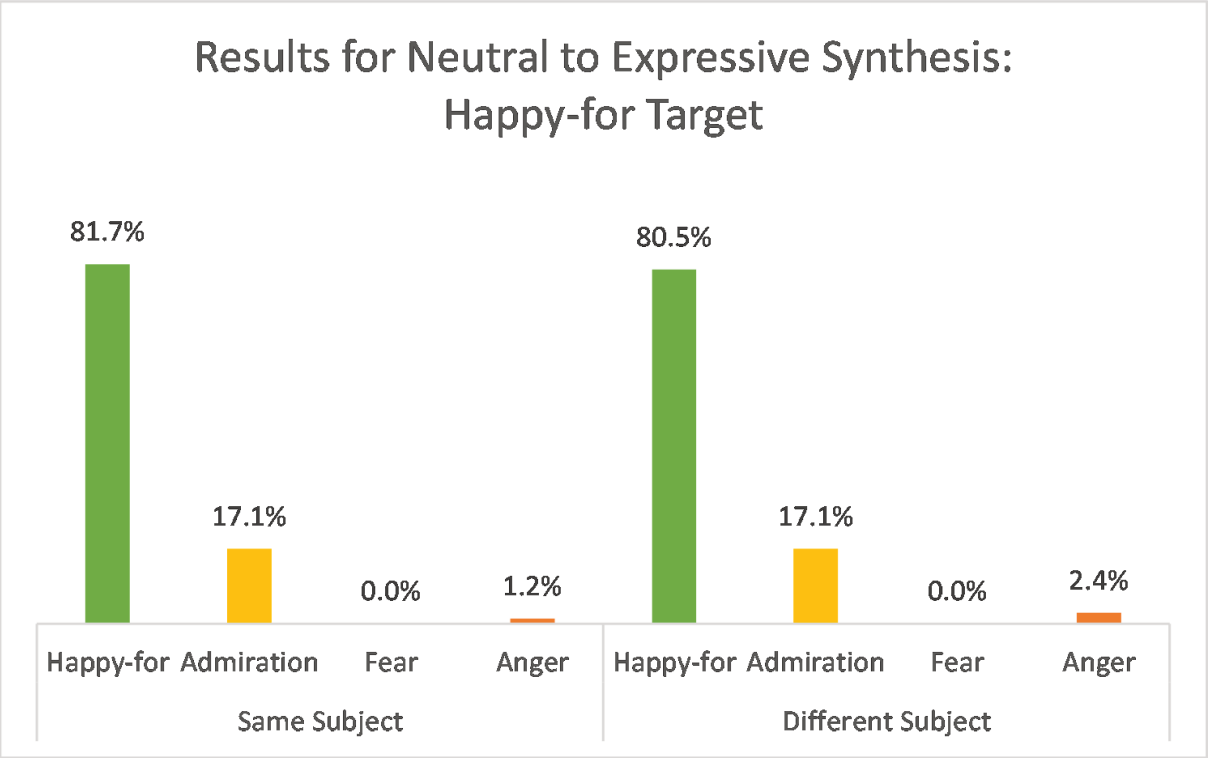


Figure 4.14 – Results for the emotion perception test with samples generated with the *Happy-for* emotion target for both **neutral to expressive synthesis - same subject**, and **neutral to expressive synthesis - different subject** stimuli. The green bars represent the occurrences when the user had chosen the correct emotion, while the yellow represents when another emotion with the same valence was chosen. The orange bars represent choices with incorrect valence. For both stimuli, more than 80 % of the choices were correct, and more than 97% of the choices represented the correct valence.

Comparing the results for *Happy-for*, and *Admiration* target, we note that the mix of choices for different emotions is more significant in the second than in the first. This more substantial choice variance demonstrates the increased challenge that representing a subtle emotion poses to the synthesis process

On the other side of the valence spectrum, *Fear* is the negative emotion closest to neutral. In Figure 4.16, we can see a clear distinction between the two stimuli groups of **neutral to expressive synthesis** test. In the stimuli set with videos generated with the same subject, the users could perceive the desired emotion on 74.4% of the instances. Additionally, the users have chosen emotions of a negative valence on 92.7% of the cases. In contrast, the target emotion was chosen on only 32.9% of the instances when considering the videos synthesized from keypoints of a different actor, and the most recurrent choice was *Anger* with 56.1% of the choices.

Finally, Figure 4.17 shows the results for the evaluation of videos synthesized with the *Anger* target. Similar to the other negative emotion, there is a clear difference between the same subject the different subject tests. In the first case, 97.6% of the choices

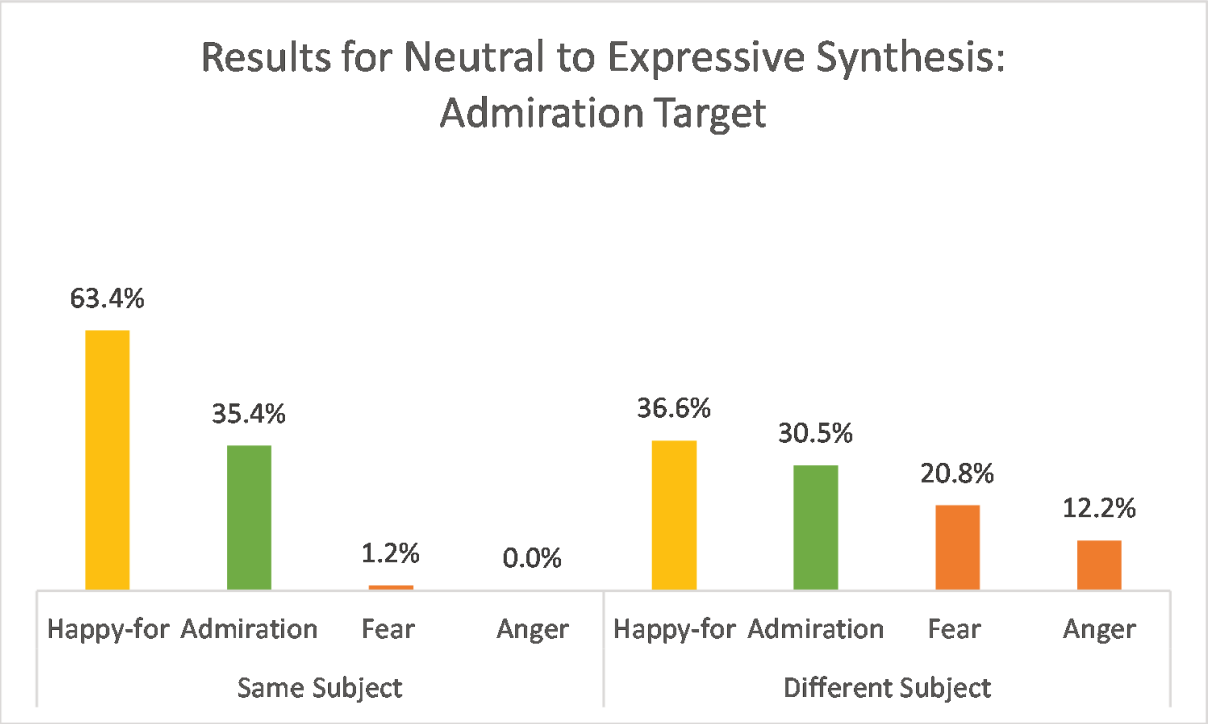


Figure 4.15 – Results for the **neutral to expressive synthesis** test with samples generated with the *Admiration* emotion target. The green bars represent the occurrences when the user had chosen the correct emotion, while the yellow represents when the users have chosen another emotion with the same valence. The orange bars represent choices with incorrect valence. In both stimuli, most of the users have chosen *Happy-for*, which is the same valence as the target emotion. For the same subject test, 33% of the responses were of incorrect valence, which shows that the users had more difficulty in perceiving the emotion and valence in this stimulus.

were correct, while in the second, 68.3% of them were right. In this case, we can also note that the spread of choices on the different subject tests is larger than on the same subject, indicating that the emotions displayed in the different subject tests are less clear to the users.

Comparing the results in the **Neutral to Expressive Synthesis** tests, we note that our system achieves the best results in the *Strong Positive* and *Strong Negative* emotions. This difference in performance is linked to the increased challenge to synthesize subtle emotions. Additionally, we note that our results are worse to the *Admiration* target. This result is led mostly by a confusion with *Happy-for* in the **same subject** results, and *Happy-for* and *Fear* in the **different subject** test.

Another approach to analyzing the emotion perception test results is to consider the valence of the target emotions and user choices. Figure 4.18 shows the results considering the valence of the target emotions and choices.

Our results indicate that our system can produce results with facial expressions

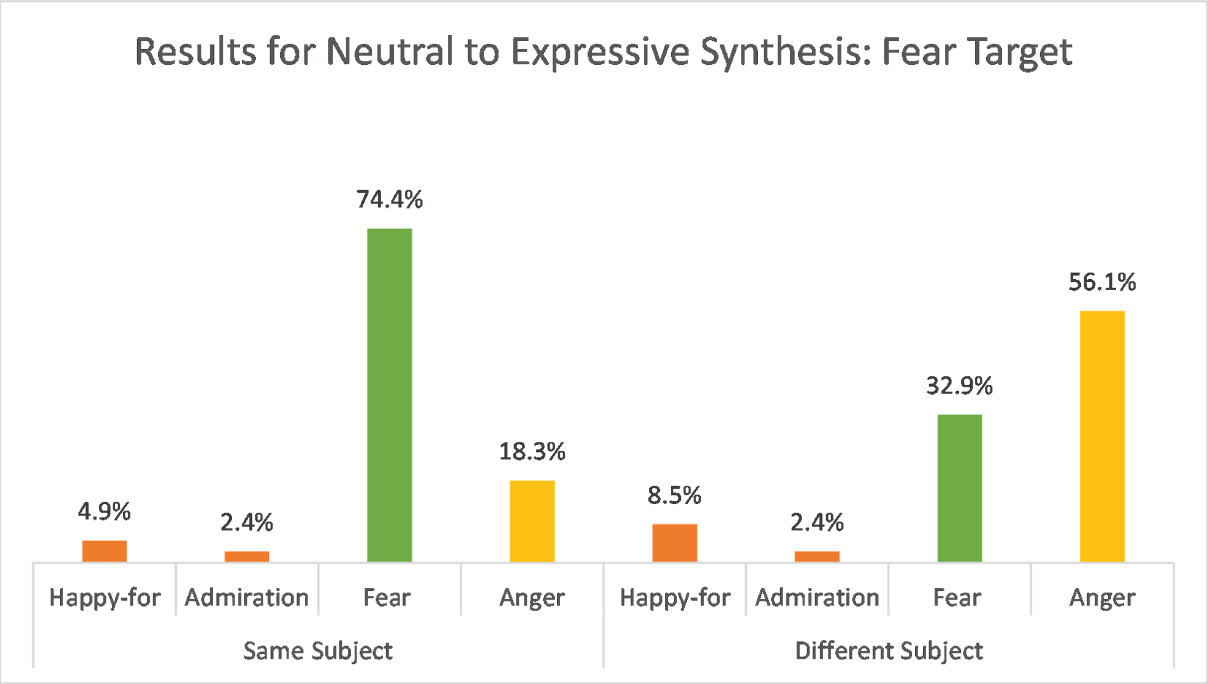


Figure 4.16 – Results for the **neutral to expressive synthesis** emotion perception test with samples generated with the *Fear* emotion target for both Stimuli groups, with keypoints from the same actress, and with keypoints from another actor. The green bars represent the occurrences when the user had chosen the correct emotion, while the yellow represents when the users chose another emotion with the same valence. In the different subject tests, 74.4% of the choices were correct, and only 7.3% of them had the wrong valence. In contrast for stimulus 2, the most common option was *Anger*, chosen on 56.1% of the cases, which relates to the challenge of synthesizing video using keypoints from an actor absent from the training dataset.

perceived as having the target valence, as the majority of users could perceive emotions within the correct valence. As expected, the same subject test results are better than the different subjects, as the first achieve success on 97% of the instances and the second on 85.4%. This result is aligned with what we presented in Section 3.4, where our approach achieved better results on objective metrics than *vid2vid*.

4.5.2 Synthesis Comparison

The results for the **synthesis method comparison** stimuli, are presented in Figure 4.19. In this test, we asked the users to choose between two videos, which one he/she considered more realistic. Our approach was chosen 54.7% of the interactions. This result shows that our methodology was capable of adding emotion expressions to the final animation without degrading, or even improving, the perception of videorealism.

Finally, we applied a binomial test to evaluate whether the preference for our approach was statistically significant, and obtained a *p-value* of $p = 0.106$. Additionally,

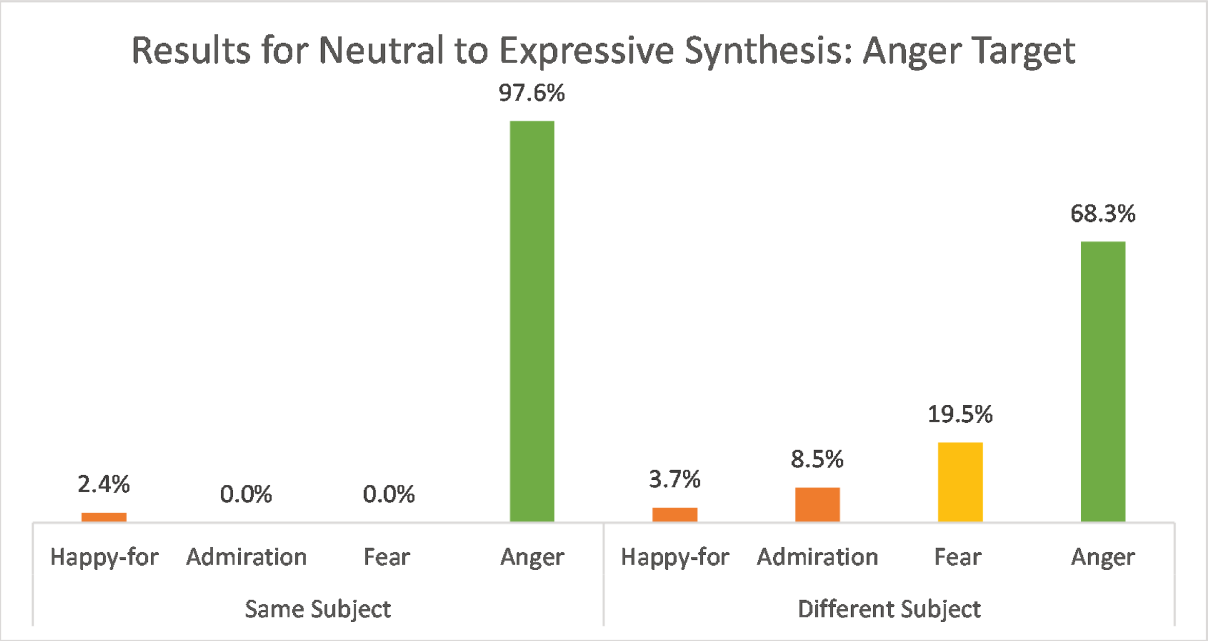


Figure 4.17 – Results for the emotion perception test with samples generated with the *Anger* emotion target for both Stimuli, with keypoints from the same subject, and with keypoints from another actor. The green bars represent the occurrences when the user had chosen the correct emotion, while the yellow represents when the users chose another emotion with the same valence. The orange bars represent choices with incorrect valence. The users chose the correct emotion in 97.6% of the instances of the same subject test, while for different subject test the choice rate was of 68.3%.

the confidence interval calculated using an asymptotic normal approximation is from 0.5 to 0.6 [3].

As the outputs of the *vid2vid* network may already be considered videorealistic by the results presented by Wang et al. (2018), the result that our system achieved in the comparison test is an excellent indicator of the great level of videorealism.

4.6 Concluding Remarks

In this chapter, we have presented the results of our videorealism assessment. We evaluated the videorealism by using a subjective test with a user group. This evaluation process is challenging, but it is essential to assess the quality of our work in a real scenario, instead of only comparing it to real videos.

The recognition of emotions through the analysis of facial expressions in a video, without audio or any other context information, is a challenging task for humans. However, our results demonstrate that even when synthesizing animations with less archetypal facial expressions, the users were able to identify the target emotion. We also observe this ability

[3] The code used in this evaluation is available at <https://github.com/fireis/masters_results_analysis>

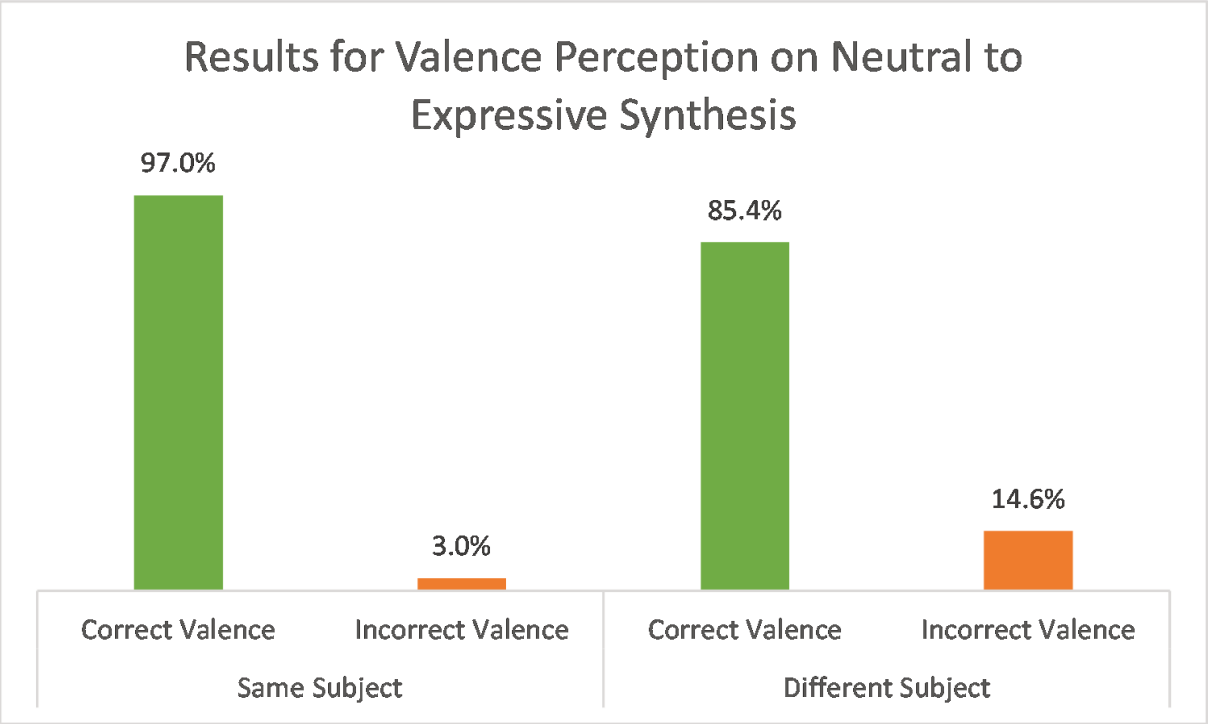


Figure 4.18 – Results for the emotion perception test considering the valence of the target emotion and the choices. The green bars represent the occurrences when the user has chosen the correct valence, while the orange ones represent choices with the incorrect valence. For both stimuli, most users chose the correct valence, which indicates that the system is capable of synthesizing videos in which the users are capable of perceiving the correct valence.

when using inputs obtained from another actor, demonstrating that our methodology showed the potential to generalize well to multiple faces.

Compared to the *vid2vid* network, our approach was preferred, demonstrating that our approach was capable of adding emotion information without degrading or improving the perception of videorealism.

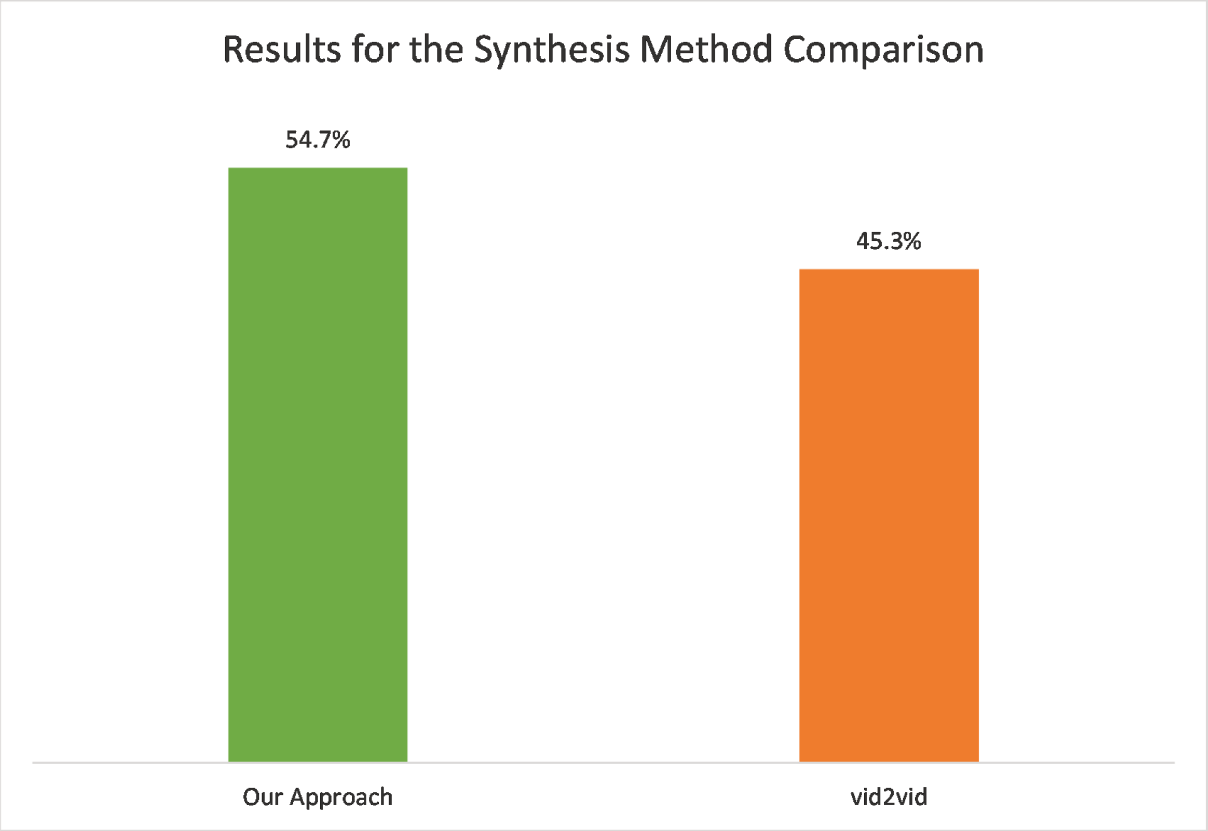


Figure 4.19 – Results for the test comparing the *vid2vid* synthesis approach to the proposed in this work. The majority of users, 54.7% of the users, chose our approach. This result indicates that our system was able to increase the perception of realism of the results. This result is also relevant to indicate that the users may associate expressive results with realism.

5 Conclusion

We proposed an expressive visual speech generation system using Generative Adversarial Networks to generate two-dimensional videorealistic results while conveying facial expressions of emotions.

In Chapter 2, we described Generative Adversarial Networks (GANs), to assure the understanding of the rest of the thesis. In this chapter, we reviewed the relevant literature for our work in the visual speech area, presenting historical contributions to the challenge of animating a videorealistic facial animation. We also presented new developments in facial animation using GANs. Finally, we introduced and discussed the emotion model used in our work.

We presented the methodology of this work in Chapter 3. In this chapter, we presented *vid2vid*, the network used as the foundation of our network. We then presented our contributions and approaches to adding the emotion information to *vid2vid*. Additionally, we presented our first results using objective metrics and qualitative evaluation. The objective metric *FID* shows that our results have a level of videorealism similar to *vid2vid*, which demonstrates that we at least maintained the videorealism after adding the emotion information. By analyzing the qualitative results, we can verify that our results are realistic and that our system allows transposing the facial keypoints from a subject to a completely different target. We showed that the system usually generalizes well to input keypoints from other subjects. However, we may produce some undesired visual artifacts when generating new videos using keypoints from a different subject as specific head movements of this subject may be completely absent from the training set, imposing difficulty on the generation system.

We presented our results in Chapter 4. Our results show that the users could associate the expressive speech synthesized by our system with the targeted emotion. This was true even when the keypoints used to synthesize the videos originated from a subject not included in the training set. Additionally, we provided results for the comparison between our approach and *vid2vid*, which results may be considered videorealistc. Our approach was usually preferred over *vid2vid*, which indicates that methodology was capable of adding emotion expressions to the final animation without degrading, or even improving, the perception of videorealism.

These results allow us to consider that our system achieves results at least as good as our reference network, *vid2vid*, which is considered to produce videorealistic results (WANG et al., 2018). Our system achieves these results in expressive video speech while generalizing well for keypoints extracted from different faces. The user evaluation

group results also demonstrate that our system can generate expressive results with the desired target emotion and valence.

5.1 Future Work

The results of our evaluation provided useful pointers on approaches to improve our method. Based on the perceptions described in Section 4.1, we know that our system has lower performance when the keypoints are extracted from a subject with different movement patterns than the subject in the training set. One approach to mitigate this effect would be to use data augmentation techniques, such as translation and rotation, to artificially increase the variance of the movements performed by the subject in the training set. Another approach would be to use videos from subjects with higher head movement variability.

Another future work approach is to incorporate our approach as the video synthesis part of a full text-to-video pipeline. This could be done by using keypoints generated by a text-to-keypoint system as inputs to our system.

Bibliography

ANDERSON, R. et al. Expressive visual text-to-speech using active appearance models. In: *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*. USA: IEEE Computer Society, 2013. (CVPR '13), p. 3382–3389. ISBN 9780769549897. Disponível em: <https://doi.org/10.1109/CVPR.2013.434>. Cited in page 32.

BREGLER, C.; COVELL, M.; SLANEY, M. Video rewrite: Driving visual speech with audio. In: *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*. USA: ACM Press/Addison-Wesley Publishing Co., 1997. (SIGGRAPH '97), p. 353–360. ISBN 0897918967. Disponível em: <https://doi.org/10.1145/258734.258880>. Cited in page 32.

BULAT, A.; TZIMIROPOULOS, G. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In: *International Conference on Computer Vision*. [S.l.: s.n.], 2017. Cited in page 48.

CHOI, Y. et al. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 8789–8797, 2017. Cited in page 35.

COSATTO, E.; GRAF, H. Photo-realistic talking-heads from image samples. *Multimedia, IEEE Transactions on*, v. 2, p. 152 – 163, 10 2000. Cited in page 32.

COSTA, P. *Two-Dimensional Expressive Speech Animation*. Tese (Doutorado) — Universidade Estadual de Campinas, 02 2015. Cited 7 times in pages 9, 32, 36, 50, 53, 57, and 61.

COSTA, P. D. P.; MARTINO, J. M. D. Assessing the visual speech perception of sampled-based talking heads. In: *AVSP*. [S.l.: s.n.], 2013. Cited in page 56.

DEY, P.; MADDOCK, S.; NICOLSON, R. Evaluation of a viseme-driven talking head. In: . [S.l.: s.n.], 2010. p. 139–142. Cited in page 56.

EKMAN, P. Facial expression and emotion. *American Psychologist*, v. 48, n. 4, p. 384–392, 1993. ISSN 1935-990X(Electronic),0003-066X(Print). Cited in page 61.

EZZAT, T.; GEIGER, G.; POGGIO, T. Trainable videorealistic speech animation. *ACM Trans. Graph.*, Association for Computing Machinery, New York, NY, USA, v. 21, n. 3, p. 388–398, jul. 2002. ISSN 0730-0301. Disponível em: <https://doi.org/10.1145/566654.566594>. Cited 2 times in pages 32 and 56.

FAN, B. et al. A deep bidirectional lstm approach for video-realistic talking head. *Multimedia Tools and Applications*, 09 2015. Cited in page 32.

FILNTISIS, P. et al. Video-realistic expressive audio-visual speech synthesis for the greek language. *Speech Communication*, v. 95, p. 137–152, 12 2017. Cited 2 times in pages 56 and 57.

FRASER, J.; PAPAIOANNOU, I.; LEMON, O. Spoken conversational ai in video games: Emotional dialogue management increases user engagement. In: *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. New York, NY, USA: Association for Computing Machinery, 2018. (IVA '18), p. 179–184. ISBN 9781450360135. Disponível em: <https://doi.org/10.1145/3267851.3267896>. Cited in page 21.

FRIED, O. et al. Text-based editing of talking-head video. *ACM Trans. Graph.*, Association for Computing Machinery, New York, NY, USA, v. 38, n. 4, jul. 2019. ISSN 0730-0301. Disponível em: <https://doi.org/10.1145/3306346.3323028>. Cited 3 times in pages 34, 56, and 57.

GOODFELLOW, I. J. Nips 2016 tutorial: Generative adversarial networks. *ArXiv*, abs/1701.00160, 2017. Cited in page 24.

GOODFELLOW, I. J. et al. Generative adversarial networks. *ArXiv*, abs/1406.2661, 2014. Cited 5 times in pages 21, 24, 26, 29, and 33.

HEUSEL, M. et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2017. (NIPS'17), p. 6629–6640. ISBN 9781510860964. Cited 4 times in pages 10, 31, 50, and 51.

HONG, Y. et al. How generative adversarial networks and their variants work: An overview. *ACM Comput. Surv.*, v. 52, p. 10:1–10:43, 2019. Cited in page 30.

HU, M.; GUO, J. Facial attribute-controlled sketch-to-image translation with generative adversarial networks. *EURASIP Journal on Image and Video Processing*, v. 2020, n. 1, p. 2, jan. 2020. ISSN 1687-5281. Disponível em: <https://doi.org/10.1186/s13640-020-0489-5>. Cited in page 34.

HUANG, H.; YU, P. S.; WANG, C. An Introduction to Image Synthesis with Generative Adversarial Nets. *arXiv:1803.04469 [cs]*, mar. 2018. ArXiv: 1803.04469. Disponível em: <http://arxiv.org/abs/1803.04469>. Cited in page 21.

ISOLA, P. et al. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017. Cited 3 times in pages 9, 33, and 41.

KARRAS, T.; LAINE, S.; AILA, T. A style-based generator architecture for generative adversarial networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2019. p. 4401–4410. Cited in page 34.

KARRAS, T. et al. Analyzing and improving the image quality of stylegan. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2020. p. 8110–8119. Cited in page 34.

Kazemi, H. et al. Facial attributes guided deep sketch-to-photo synthesis. In: *2018 IEEE Winter Applications of Computer Vision Workshops (WACVW)*. [S.l.: s.n.], 2018. p. 1–8. ISSN null. Cited in page 34.

KIM, H. et al. Deep video portraits. *ACM Trans. Graph.*, v. 37, p. 163:1–163:14, 2018. Cited 2 times in pages 56 and 57.

Kim, K. et al. Effects of patient care assistant embodiment and computer mediation on user experience. In: *2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. [S.l.: s.n.], 2019. p. 17–177. ISSN null. Cited in page 20.

KING, D. E. Dlib-ml: A machine learning toolkit. *J. Mach. Learn. Res.*, JMLR.org, v. 10, p. 1755–1758, dez. 2009. ISSN 1532-4435. Disponível em: <http://dl.acm.org/citation.cfm?id=1577069.1755843>. Cited in page 48.

LI, C.; WAND, M. Precomputed real-time texture synthesis with markovian generative adversarial networks. In: LEIBE, B. et al. (Ed.). *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III*. Springer, 2016. (Lecture Notes in Computer Science, v. 9907), p. 702–716. Disponível em: <https://doi.org/10.1007/978-3-319-46487-9_43>. Cited in page 33.

Liu, K.; Ostermann, J. Realistic facial expression synthesis for an image-based talking head. In: *2011 IEEE International Conference on Multimedia and Expo*. [S.l.: s.n.], 2011. p. 1–6. ISSN 1945-7871. Cited in page 32.

LUčIć, M. et al. Are gans created equal? a large-scale study. In: *Advances in Neural Information Processing Systems*. [s.n.], 2018. Disponível em: <https://arxiv.org/pdf/1711.10337.pdf>. Cited in page 31.

MARTINO, J. M. D.; MAGALHãES, L. P.; VIOLARO, F. Facial animation based on context-dependent visemes. *Computers & Graphics*, v. 30, n. 6, p. 971–980, dez. 2006. ISSN 0097-8493. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0097849306001518>. Cited in page 35.

MATTHEYSES, W.; VERHELST, W. Audiovisual speech synthesis: An overview of the state-of-the-art. *Speech Communication*, v. 66, p. 182–217, fev. 2015. ISSN 0167-6393. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0167639314000818>. Cited 6 times in pages 20, 21, 31, 32, 35, and 56.

MIRZA, M.; OSINDERO, S. Conditional generative adversarial nets. *ArXiv*, abs/1411.1784, 2014. Cited 2 times in pages 29 and 33.

ORTONY, A.; CLORE, G.; COLLINS, A. *The Cognitive Structure of Emotion*. [S.l.: s.n.], 1988. v. 18. Cited in page 61.

PANDZIC I., O. J. M. D. *The Visual Computer - Springer*. 1985. Disponível em: <https://link.springer.com/journal/371>. Cited in page 20.

PARKE, F. I. Computer generated animation of faces. In: *Proceedings of the ACM Annual Conference - Volume 1*. New York, NY, USA: Association for Computing Machinery, 1972. (ACM '72), p. 451–457. ISBN 9781450374910. Disponível em: <https://doi.org/10.1145/800193.569955>. Cited in page 32.

QIAO, F. et al. Emotional facial expression transfer from a single image via generative adversarial nets. *Computer Animation and Virtual Worlds*, v. 29, n. 3-4, p. e1819, 2018. Cited 3 times in pages 34, 35, and 56.

RADFORD, A.; METZ, L.; CHINTALA, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015. Cited 2 times in pages 30 and 33.

RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. In: NAVAB, N. et al. (Ed.). *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Cham: Springer International Publishing, 2015. p. 234–241. ISBN 978-3-319-24574-4. Cited in page 33.

RUHLAND, K.; PRASAD, M.; MCDONNELL, R. Data-driven approach to synthesizing facial animation using motion capture. *IEEE Computer Graphics and Applications*, v. 37, n. 4, p. 30–41, 2017. Cited 2 times in pages 56 and 57.

SALIMANS, T. et al. Improved techniques for training gans. In: LEE, D. D. et al. (Ed.). *Advances in Neural Information Processing Systems 29*. Curran Associates, Inc., 2016. p. 2234–2242. Disponível em: <http://papers.nips.cc/paper/6125-improved-techniques-for-training-gans.pdf>. Cited in page 31.

SUWAJANAKORN, S.; SEITZ, S. M.; KEMELMACHER-SHLIZERMAN, I. Synthesizing obama: Learning lip sync from audio. *ACM Trans. Graph.*, Association for Computing Machinery, New York, NY, USA, v. 36, n. 4, jul. 2017. ISSN 0730-0301. Disponível em: <https://doi.org/10.1145/3072959.3073640>. Cited in page 34.

Szegedy, C. et al. Rethinking the inception architecture for computer vision. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2016. p. 2818–2826. Cited in page 30.

VONDRICK, C.; PIRSIAVASH, H.; TORRALBA, A. Generating videos with scene dynamics. In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2016. (NIPS'16), p. 613–621. ISBN 9781510838819. Cited in page 33.

WALKER, J. H.; SPROULL, L.; SUBRAMANI, R. Using a Human Face in an Interface. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 1994. (CHI '94), p. 85–91. ISBN 978-0-89791-650-9. Event-place: Boston, Massachusetts, USA. Disponível em: <http://doi.acm.org/10.1145/191666.191708>. Cited in page 20.

WANG, T.-C. et al. Few-shot video-to-video synthesis. In: *Advances in Neural Information Processing Systems (NeurIPS)*. [S.l.: s.n.], 2019. Cited 3 times in pages 31, 34, and 56.

WANG, T.-C. et al. Video-to-video synthesis. In: *Conference on Neural Information Processing Systems (NeurIPS)*. [S.l.: s.n.], 2018. Cited 12 times in pages 31, 33, 37, 39, 40, 41, 44, 49, 56, 57, 73, and 76.

WANG, W. et al. Comp-gan: Compositional generative adversarial network in synthesizing and recognizing facial expression. In: *Proceedings of the 27th ACM International Conference on Multimedia*. New York, NY, USA: Association for Computing Machinery, 2019. (MM '19), p. 211–219. ISBN 9781450368896. Disponível em: <https://doi.org/10.1145/3343031.3351032>. Cited in page 34.

XIE, L.; SUN, N.; FAN, B. A statistical parametric approach to video-realistic text-driven talking avatar. *Multimedia Tools Appl.*, Kluwer Academic Publishers, USA, v. 73, n. 1, p. 377–396, nov. 2014. ISSN 1380-7501. Disponível em: <https://doi.org/10.1007/s11042-013-1633-3>. Cited in page 32.

ZAKHAROV, E. et al. Few-shot adversarial learning of realistic neural talking head models. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Oct 2019. Disponível em: <http://dx.doi.org/10.1109/ICCV.2019.00955>. Cited 3 times in pages 31, 34, and 56.

ZHU, J.-Y. et al. Unpaired image-to-image translation using cycle-consistent adversarial networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, p. 2242–2251, 2017. Cited in page 33.