Generating Facial Expression Sequences of Complex Emotions with Generative Adversarial Networks

Zakariae Belmekki, David Gomez, Patrick Reuter, Jun Li, Jean-Claude Martin, Karl W. Jenkins, Nadine Couture

Generating Facial Expression Sequences of Complex Emotions with Generative Adversarial Networks

Zakariae Belmekki
Univ. Bordeaux, ESTIA-Institute of Technology, EstiaR, F-64210 Bidart
France
Centre for Computational Engineering Sciences , Cranfield University
United Kingdom
zakariae.belmekki@estia.fr

David Antonio Gómez Jáuregui
Univ. Bordeaux, ESTIA-Institute of Technology, EstiaR, F-64210 Bidart
France
d.gomez@estia.fr

Patrick Reuter
Univ. Bordeaux, Inria, Bordeaux INP, CNRS (LaBRI UMR 5800), ESTIA, Bordeaux
France
preuter@labri.fr

Jun Li
Centre for Computational Engineering Sciences, Cranfield University
United Kingdom
jun.li@cranfield.ac.uk

Jean-Claude Martin
Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, 91400, Orsay
France
jean-claude.martin@universite-paris-saclay.fr

Karl W Jenkins
Centre for Computational Engineering Sciences, Cranfield University
United Kingdom
k.w.jenkins@cranfield.ac.uk

Nadine Couture
Estia-Recherche, Univ. Bordeaux, ESTIA-Institute of Technology, EstiaR, F-64210 Bidart
France
n.couture@estia.fr

Abstract

There is a rising interest in animating realistic virtual agents for multiple purposes in different domains. Such a task requires systems capable of generating complex mental states on par with human emotional complexity. Considering the high representational capacity of Generative Adversarial Networks (GANs), it is only natural to consider them in such applications. In this work, we propose a conditional GAN model for generating sequences of facial expressions of categorical complex emotions. Trained on a scarce and highly imbalanced dataset, the proposed model is able to generate realistic variable-length sequences in a single inference step. These expressions of emotional states, of which there are 24 in total, follow the Facial Actions Coding System (FACS) formatting. In the absence of meaningful objective evaluation methods, we propose a deep-learning-based metric to assess the realism of generated Action Unit (AU) sequences: the Action Unit Fréchet Inception Distance (AUFID). Objective and subjective results validate the realism of our generated samples.

ICMI '24, November 04–08, 2024, San Jose, Costa Rica

ACM ISBN 979-8-4007-0462-8/24/11
https://doi.org/10.1145/3678957.3685712

CCS Concepts

• **Computing Methodologies** → **Machine Learning**.

Keywords

Generative Adversarial Networks, Complex Emotional States, Synthetic Action Units, FACS

ACM Reference Format:
Zakariae Belmekki, David Antonio Gómez Jáuregui, Patrick Reuter, Jun Li, Jean-Claude Martin, Karl W Jenkins, and Nadine Couture. 2024. Generating Facial Expression Sequences of Complex Emotions with Generative Adversarial Networks. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '24), November 04–08, 2024, San Jose, Costa Rica.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3678957.3685712

1 Introduction

Virtual agents can be used in multiple domains, be it medical or even digital entertainment. Building realistic virtual agents is a field of growing interest. An important factor to consider to improve the realism of virtual agents is their affective capabilities, that is, the ability to manifest emotional states that are on par with human emotion complexity. Several applications of socially interactive agents require complex mental states that go beyond the six basic emotional states (sad, happy, angry, afraid, surprised, disgusted) [22]. For this reason, there is a rising interest in generating facial expressions of complex mental states. Several works and experimental studies showed the advantage of generating such data for affective interaction with virtual agents [9, 26]. However, the main difficulty

lies in the complex dynamics that characterize facial expressions of emotions. Generating them requires a system capable of reproducing the high level of complexity present in natural human mental states. Simplistic methods, like linear interpolation across all areas of the face between static facial expressions, lead to non-realistic animations [1, 29]. Therefore, Deep Learning models appear as a compelling solution to the aforementioned challenge. In particular, Generative Adversarial Networks (GANs) [15] are a good candidate for such generative tasks.

On the one hand, existing works [11, 12] that used GAN to generate sequences of facial expressions are conditioned on speech, and therefore cannot produce distinct categorical complex emotional expressions, which is a different application from co-speech generation. Additionally, such applications of generative models suffer from a significant challenge: the absence of meaningful objective evaluation metrics. The metrics used in relevant previous works do not capture meaningful salient characteristics of sequential AU data [12]. Having access to a reliable objective metric would make the process of developing and fine-tuning generative models for such applications cheaper and considerably faster. On the other hand, the existing literature for generating facial expressions of emotional states [18, 28] is limited to the 6 basic emotions and their proposed methods can only generate static outputs, i.e., single frames or images with no time component.

An additional important challenge in this domain is the scarcity of datasets of labeled sequences of facial expressions of complex emotions [27]. This hinders the adoption of deep-learning-based methods in applications like ours since Deep Learning models require large numbers of training samples.

In this context, the present work answers the aforementioned limitations by proposing the following:

(1) A new GAN-based pipeline for generating realistic AU sequences of facial expressions representing complex emotional states in the FACS format [14], trained on a highly imbalanced and scarce dataset.
(2) Our GAN model generates variable-length sequences in one inference step, in contrast to the commonly used method of generating samples at multiple regular intervals [12], where each sample is generated with multiple inference steps.
(3) A deep-learning-based objective method is proposed to evaluate the realism of generated sequences of Action Units (AUs), which we call the Action Unit Fréchet Inception Distance (AUFID).

Additionally, objective and subjective evaluations of our results show that our model successfully generates realistic AU sequences.

2 Related Work

Most of the existing literature is limited to single frame/image generation. [30] proposed GANimation for conditioning faces on AUs. While their generated samples were realistic, fine-grained changes to facial expressions were challenging to perform. This was alleviated by two works: [20] proposed LAC-GAN, characterized with a local attention mechanism, it could change an AU without affecting others. Similarly, [19] proposed a method for fine-grained facial expression editing using relative AUs. Similar to [30], [31] propose PattGAN, which is based on StarGAN [8], to condition images on AUs. Other works have adopted 3D intermediary representations to generate single-frame facial expressions. [21] trained their GAN model on 3D Morphable Models (3DMM) representations extracted from images and conditioned on AUs. Similarly, [27] trained their GAN model on rasterized 3D meshes (single 2D frames). None of the aforementioned works generated sequences of facial expressions, they all generate single frames. The problem of generating sequences is a more challenging one, as it requires the model to learn not only AU representations but also their dynamics across the temporal axis.

As for generating facial expressions of categorical emotions, the existing works use only basic emotions. [18] builds on GANimation by adding another conditional GAN to condition the generated images on discrete categorical emotions that are: anger, neutral, fear, happiness, sadness and surprise. Also, this work belongs to the previously described category that only generates single frames. [28] proposed generating motions of six basic emotions, given a neutral face image. The generation, however, is done using facial landmarks and not AUs. Here lies another contribution of the present work: our model generates 24 facial expressions of complex emotions, instead of the basic 6.

While few existing works generate facial expression sequences, none of them apply it in a complex emotion context. [7] used a GAN with a Recurrent Neural Network (RNN) architecture to generate facial expressions, but they are only limited to lip animation for static face images. The AUs involved are speech-related ones. Most relevant to our work are [12] and [11], where GAN architectures are proposed to generate head and gaze movements with facial expressions. The two main differences between those works and the present work are the following: (1) their application is co-speech facial movements generation, in contrast to ours, which is generating facial expressions of complex emotions, and (2) the architectures are different in that our model generates variable-length AU sequences, while theirs can only generate segments of 4 seconds. This implies that their model does not learn the continuity of features in the input speech sequence beyond 4 seconds.

3 Methodology

3.1 Dataset and Preprocessing

The model is trained on two merged datasets: MindReading [4], which contains 2471 videos of 24 categorical emotional states, and Padova Emotional Dataset of Facial Expressions (PEDFE) [24], which contains 1734 videos of 6 categorical emotional states (see Figure 2). The 6 emotional states in PEDFE are also present in MindReading, allowing the merging of the two datasets. The resulting dataset consists of 4205 samples with a median length of 5 seconds, of which we keep 3659 after removing sequences with less than 40 frames and more than 300 frames, which are outliers. These sequences consist of frames captured every 0.04 seconds. The dataset is not only scarce but also suffers from severe data imbalance, which makes training a Deep Learning model challenging. However, succeeding in training a GAN in this setting would be a strong indicator of the effectiveness of GAN for generating facial expressions for affective interaction purposes, where data is often lacking.

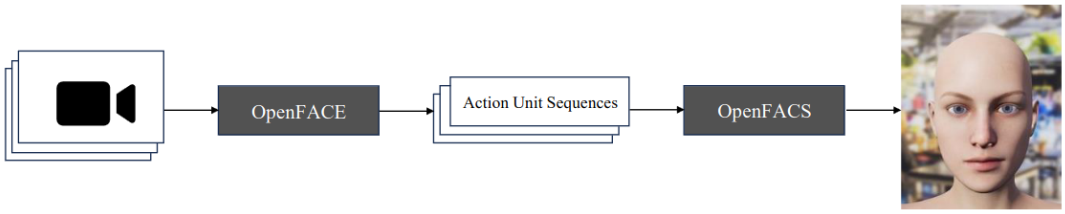


Figure 1: The pipeline for extracting and visualizing sequences of emotional states. Videos are processed by OpenFACE to extract AUs, which can then be visualized on OpenFACS.

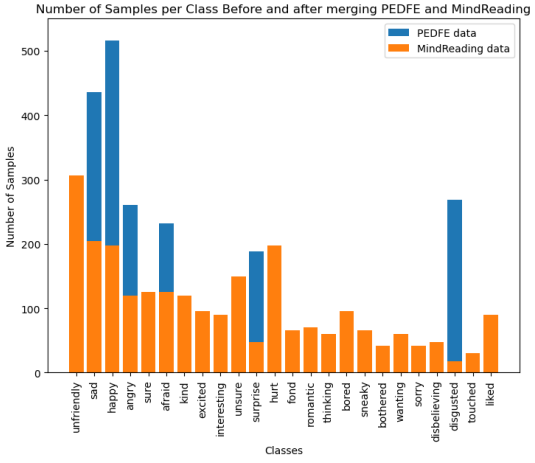


Figure 2: The distributions of the 24 categorical emotional states in the dataset. The number of samples from the MindReading dataset is represented in orange, and the number of samples from the PEDFE dataset is represented in blue.

The pipeline, presented in Figure 1, is as follows: videos from the merged datasets are processed by OpenFACE [3] to extract sequences of 17 Action Unit (AU) intensities. Since the videos are of different lengths, the sequences extracted by OpenFACE vary in size. This is problematic since the inputs of our model must be of the same size. To remedy this, we set a maximum length size of 300 for sequences and add zero padding to samples of length smaller than 300. This results in samples of shape (300×17). For the final visualization of emotional states, we use OpenFACS [10].

Note that the AU intensities of the sequences generated by OpenFACE range from 0 to 5, but we scale them to the range $[-1, 1]$. The reason behind this is that the output of our GAN is in the range $[-1, 1]$ as well, since the output activation is Tanh (see Section 3.3).

3.2 Sample Format and Visualisation

We use a qualitative visualization method to monitor the quality of generated samples during and after training, as it is not convenient to visualize them on OpenFACS at every epoch. Also, as explained in Section 4.1, there are intricate features that cannot be seen when samples are visualized on the avatar.

Data samples are matrices of size (300×17), where 300 is the maximum number of frames and 17 is the number of AU categories. Every element in the matrix corresponds to an AU intensity at a specific frame. Therefore, visualizing them as heat maps makes sense, where one axis would correspond to the AU category and the second to the frame number. Additionally, in order to have an overall visual of the variations of AU intensities across time, one can visualize the mean AU intensity along the AU category axis. Compared to the heat map, the curve of mean AU intensities shows more fine-grained details of the sequence across time. Figure 3 shows a real sequence visualized using the described method.

Visualizing samples in this manner allows one to capture the following three characteristics of real facial expression AU sequences:

(1) The blue curve of the mean AU intensities exhibits subtle variations over the temporal axis, in contrast to samples generated by an untrained GAN, which are characterized by a smooth curve (see Figures 6b and 6c).
(2) The AU intensities are all zero after a certain number of frames, due to the zero padding. The number of frames after which a sequence becomes zero should vary in the generated samples since the original sequences vary in length as well.
(3) The activations of AU intensities tend to form continuous lines across time in the heat map. This is due to the nature of facial movements, which are finite and continuous in time.

The presence of these three aspects is an indicator of successful GAN training, although they do not capture the semantic aspect of samples. This, however, is compensated by the use of Principle Component Analysis (PCA) presented in Section 4.1, where the clusters formed of real and generated data points for every emotional state are compared.

3.3 GAN Architecture

GAN consists of two models: a Generator and a Discriminator, denoted G and D, respectively [15]. Both models train in an adversarial manner, wherein the Generator is trained to generate samples resembling the training data to trick the Discriminator into classifying them as real data samples, and the Discriminator is trained to correctly classify real and generated samples. The original loss function is as follows

$$\min_G \max_D \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))], \quad (1)$$

where $p_{data}(x)$ and $p_z(z)$ denote the real data probability distribution and a prior from which random noise is sampled, respectively. As the equation suggests, the Generator generates samples from a random input noise from $p_z(z)$: it learns a mapping from a prior to a probability distribution approximating the real data distribution $p_{data}(x)$. In this work, the Least Squares GAN (LSGAN) loss function [23] is adopted instead of the vanilla one, as the LSGAN loss ensures a more stable training and leads to higher quality samples [23]. Additionally, the proposed model is a conditional GAN

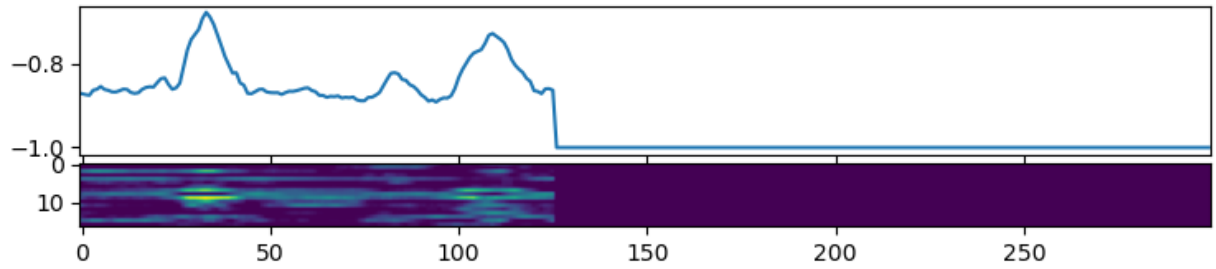


Figure 3: AU sequence visualization. The lower figure is a heat map of a transposed sequence of size (300×17) and the upper blue curve is that of the mean AU intensities along the AU category dimension.

[25] since the output is conditioned on an input variable y that determines the category of the emotional state to be generated (amongst the previously introduced 24 categories). Therefore, the loss function is as follows

$$\min_D \frac{1}{2}\mathbb{E}_{x\sim p_{data}(x)}[(D(x|y)-1)^2] + \frac{1}{2}\mathbb{E}_{z\sim p_z(z)}[(D(G(z|y)))^2] \tag{2}$$

$$\min_G \frac{1}{2}\mathbb{E}_{z\sim p_z(z)}[(D(G(z|y))-1)^2]. \tag{3}$$

Despite the training data being time-series, considering the presence of a temporal axis, a CNN-based GAN model is adopted. GANs are already known for being hard to train, and thus, using an RNN architecture would make training even harder due to their instabilities [11]. While the data samples are multivariate time-series, we use 2D convolutions due to the existence of relationships between AUs for each emotion (any given emotion can be characterized with the activation of a set of AUs to variant degrees), hence the existence of local relationships between AUs. This choice is further motivated by an interest in exploring 2D convolutional architectures in the context of generating variable-length multivariate time series. In this case, the first convolutional layers learn local dependencies between AUs across time, while the deeper layers learn global dependencies. The architectures of the Generator and the Discriminator are presented in Tables 1. In the Generator, we use the GELU [16] activation function to ensure a better flow of gradients. It was not used in the Discriminator to increase balance during training: the Discriminator, in our setting, tends to learn quicker than the Generator, which leads to weak gradient updates for the Generator after a certain number of epochs. Generally, optimal GAN training requires both models to learn in a sort of balance where no model significantly outperforms the other. It must be noted that our architecture, in contrast to relevant existing works [11, 12], can generate variable-length sequences in one inference step, instead of generating a sequence over multiple regular intervals.

3.4 Quantitative Evaluation: Action Unit Fréchet Inception Distance (AUFID)

We propose the AUFID method to evaluate the realism of the generated samples. It is based on the existing Fréchet Inception Distance (FID) [17]. The latter uses the InceptionV3 model to extract features from both real and synthetic data samples and compare them using the Fréchet Distance. For two gaussian distributions with means and covariances (μ_1, Σ_1) and (μ_2, Σ_2), the Fréchet Distance is defined by:

$$\|\mu_1 - \mu_2\|_2^2 + Tr(\Sigma_1 + \Sigma_2 - 2(\Sigma_1\Sigma_2)^{\frac{1}{2}}) \tag{4}$$

However, applying the FID in our setting is impossible for two reasons:

(1) The InceptionV3 model was trained on the ImageNet [13] dataset, and is therefore only capable of detecting features belonging to the categories of objects present in ImageNet, and not facial expressions in the form of AUs.

Table 1: The architectures of the Generator and the Discriminator, with ConvTranspose2d(input channels, output channels, kernel size, padding), Conv2d(input channels, output channels, kernel size, padding), BatchNorm2d(output channels), BilinearUpsampling(output width size, output height size) and AvgPool2d(kernel size)

Generator
ConvTranspose2d(88, 64, (4, 2), (0, 0))
BatchNorm2d(64)
LeakyReLU(0.2)
Conv2d(64, 128, (5, 3), (4, 2))
BatchNorm2d(128)
GELU()
BilinearUpsampling(16, 6)
Conv2d(128, 256, (5, 3), (4, 2))
BatchNorm2d(256)
GELU()
BilinearUpsampling(32, 8)
Conv2d(256, 384, (5, 3), (4, 2))
BatchNorm2d(384)
GELU()
BilinearUpsampling(64, 10)
Conv2d(384, 512, (5, 3), (4, 2))
BatchNorm2d(512)
GELU()
BilinearUpsampling(128, 12)
Conv2d(512, 640, (5, 3), (4, 2))
BatchNorm2d(640)
GELU()
BilinearUpsampling(256, 14)
Conv2d(640, 640, (5, 3), (4, 2))
BatchNorm2d(640)
GELU()
BilinearUpsampling(300, 17)
Conv2d(640, 1, (1, 1), (0, 0))
Tanh()

Discriminator
Conv2d(25, 512, (5, 3), (4, 2))
BatchNorm2d(512)
LeakyReLU(0.2)
AvgPool2d(2)
Conv2d(512, 256, (5, 3), (4, 2))
BatchNorm2d(256)
LeakyReLU(0.2)
AvgPool2d(2)
Conv2d(256, 128, (5, 3), (4, 2))
BatchNorm2d(128)
LeakyReLU(0.2)
AvgPool2d(2)
Conv2d(128, 64, (5, 3), (4, 2))
BatchNorm2d(64)
LeakyReLU(0.2)
AvgPool2d(2)
Linear(2816, 1)
Sigmoid()

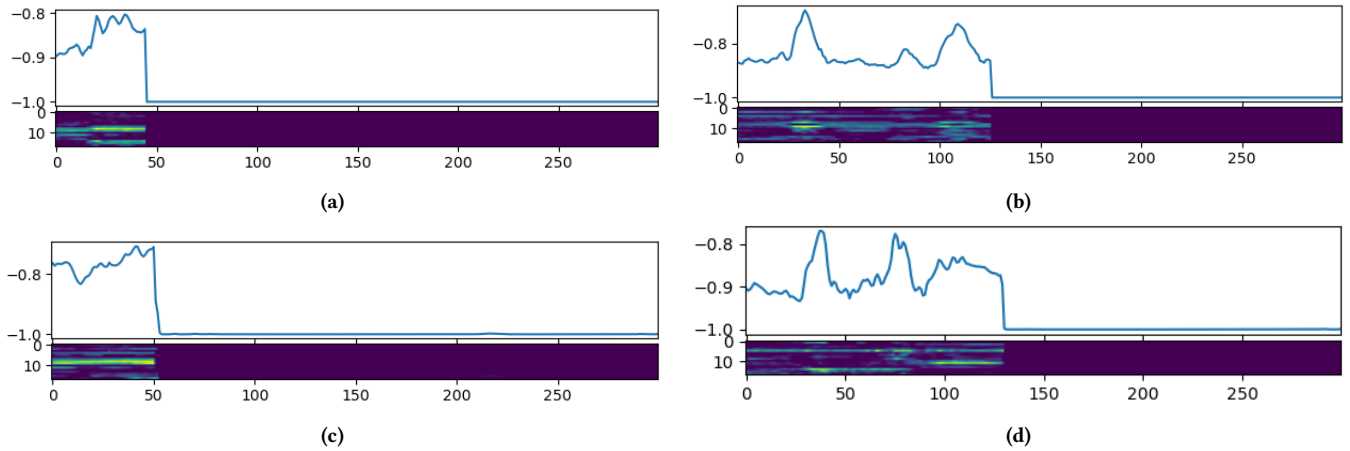


Figure 4: Two real samples ((a) and (b)) and two synthetic samples ((c) and (d)). All the samples belong to the emotional state of "surprise". The mean AU intensities curve shows the same level of variation across time and the heatmaps show relatively similar AU activations.

(2) The InceptionV3 model cannot process AU intensity sequences of size (300 × 17). It was made for RGB images of a different size.

In order to adapt the FID to our context, one has to build a model adapted to the format of data at hand and train it to recognize its features. Therefore, we build a new classifier inspired by the InceptionV3 model, adapted to process (300×17) data. Additionally, to be able to extract features from the AU intensity sequences, the classifier had to be trained on our dataset. It must be noted that, since AUFID is a distance, **the lower its value is, the more realistic the input samples are**.

3.5 Subjective Evaluation

With the help of human participants, the objective of this protocol is to verify if the Generator can successfully learn a probability distribution that approximates $p_{data}(x)$. More practically, denoting P_{real} the probability that **real samples are classified as real by human participants** and $P_{synthetic}$ the probability that **synthetic samples are classified as real by human participants**, the protocol checks how close P_{real} and $P_{synthetic}$ are. Closer values indicate smaller differences in human participants' perceptions of real and generated samples.

From the 24 emotional states, we select 5 that are the most visually recognizable by the naked eye when visualized on the OpenFACS avatar, which is also reflected by their higher mean total AU intensity. We also picked emotional states that are semantically different to avoid confusing the participants (e.g., "happy" and "excited" can be expressed with similar facial expressions, and, therefore, it is not ideal to use them in the test). The selected mental states with their corresponding mean total AU intensity are "disgusted" (8.01), "happy" (10.32), "romantic" (7.13), "touched" (8.04), and "unsure" (5.56).

Participants were shown 3 sequences of facial expressions visualized on OpenFACS for each emotional state. Two sequences are real and one is fake. The first real sequence is shown as a reference so that the participants have an idea of what a real sequence would look like on OpenFACS. Naturally, the participants are told that the reference video is a real sequence. The remaining two sequences (one is real and the other is fake) are shown to the participants, who are then asked to tell if the sequences are real or synthetic. Thus, each participant is shown 15 sequences, among which they are asked to classify 10. Considering there are 11 participants, there were 55 attempts at classifying real sequences (11 participants × 5 real sequences) and 55 attempts at classifying synthetic sequences (11 participants × 5 synthetic sequences).

The videos of facial expressions ranged from 5 to 19 seconds. Note that, with 300 frames, the maximum length of the videos should be 12 seconds. However, OpenFACS comes with a controllable speed parameter and a delay between frames was added to make sure the expressions of the avatar flow as naturally as possible. Additionally, the recording of the videos was done manually, which led to the addition of some seconds that correspond to the delay between pressing the record button and the start of the expression.

4 Experiments and Results

The model is trained for 1500 epochs using the Adam optimizer with a learning rate of 0.0002 and momentums $\beta_1 = 0.5, \beta_2 = 0.999$ for both the Generator and the Discriminator. Every 10 epochs, the model is saved. This allows us to pick the best-performing model out of all the checkpoints. The best Generator in this setting, which was saved at epoch 570, has an AUFID score of 10.48. The results show that our model could generate realistic samples with variable lengths. Using the visualization method presented in Section 3.2, we show 2 real and 2 synthetic samples for the emotional state "surprise" in Figure 4.

4.1 Objective Validation

The AUFID proved useful in providing a meaningful quantitative measure of GANs' performance, which could be used for model selection and fine-tuning. The AUFID curve is presented in Figure

5. In the final epochs, the AUFID values increased, indicating a certain level of instability. This phenomenon does not fall out of the norm, as GANs are known for their instability [2, 6]. This can be alleviated in a simple way: by saving the model with the best AUFID, similar to how early stopping works.

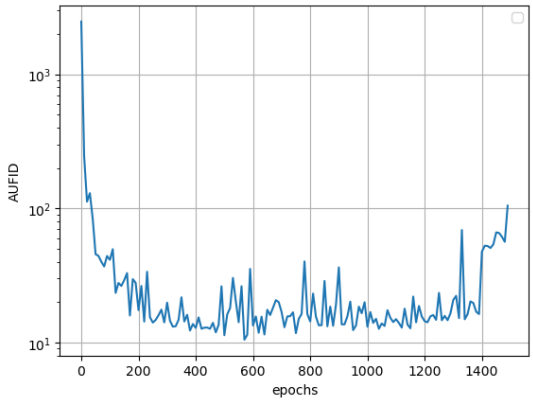
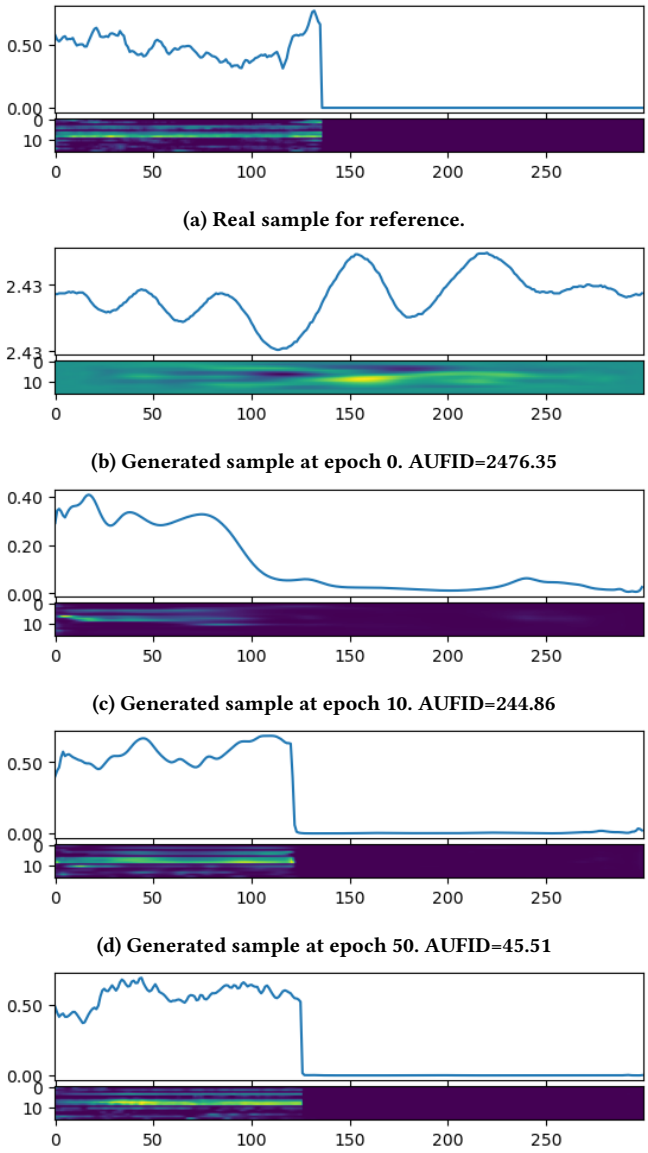


Figure 5: AUFID curve during GAN model training. The values were recorded every 10 epochs. The metric decreases throughout training, indicating ongoing successful training.

Additionally, the AUFID metric reflected well the aspects of realism presented in Section 3.2. Figure 6 shows the evolution of generated samples from the start of training (epoch 0) to the best-recorded stage of training (epoch 570, characterized with the lowest AUFID). Figure 6a is shown for reference. The sample shown in Figure 6b, generated at epoch 0, shows no resemblance to real samples (none of the 3 characteristics described in Section 3.2 are present). Naturally, it corresponds to the highest AUFID. At epoch 10, the sample 6c started exhibiting the real data characteristics to some degree: the right tail of the blue curve of the mean AU intensities started tilting toward zero. One can also observe that AU intensities in the heat map start forming subtle continuous lines along the temporal axis. At epoch 50, presented in Figure 6d, the characteristics of real samples are present to a stronger degree. However, there is some noise in the right tail of the mean AU intensities curve, where it is supposed to be constantly zero after a certain number of frames. Additionally, while the mean AU intensities blue curve shows some level of variation across time, it is still smoother than the typical real sequence (see Figure 6a for reference). Finally, at epoch 570, presented in Figure 6e, the generated sample is indistinguishable from the real sample on the three aspects considered in Section 3.2.

Note that the difference in realism between the sequences from Figures 6d and 6e is not clear when both samples are visualized on OpenFACS, as they are not very distinguishable. However, both the visualization method used in this section and the AUFID reflect well the difference in realism between the two. AUFID in particular offers an objective way of capturing the difference, assigning a lower value to the more realistic one.

While the AUFID is a good indicator of the overall sample quality, it does not offer a direct way of evaluating GANs' conditioning performance, i.e., the ability to faithfully produce samples belonging to



(a) Real sample for reference.

(b) Generated sample at epoch 0. AUFID=2476.35

(c) Generated sample at epoch 10. AUFID=244.86

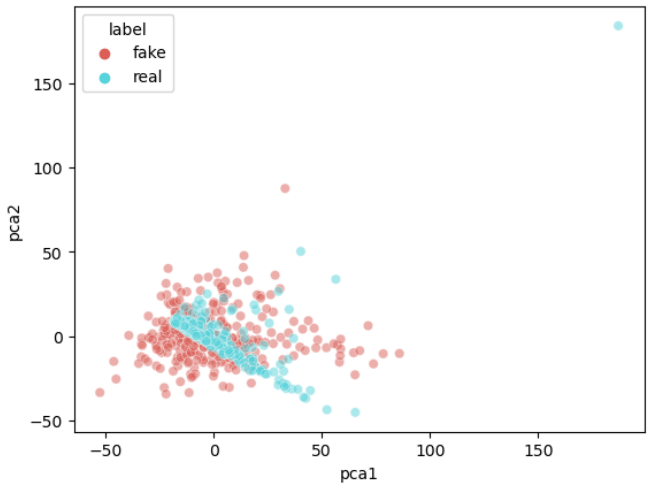
(d) Generated sample at epoch 50. AUFID=45.51

(e) Generated sample at epoch 570. AUFID=10.48

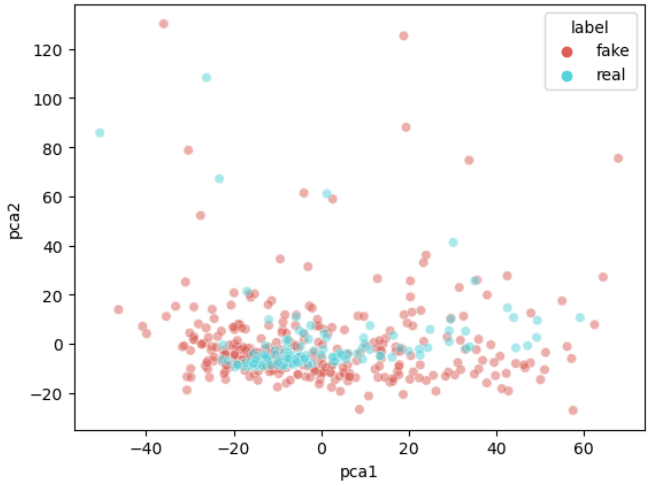
Figure 6: Synthetic samples generated at epochs 0, 10, 50, and 570 with their corresponding AUFID scores. As AUFID decreases, sample realism increases. AU intensities were scaled to the original range of $[0, 5]$.

a given emotional state category. To evaluate the conditioning performance of the model, PCA is used to visualize real and synthetic samples as two-dimensional data points for each emotional state. The benefits of using this method are twofold: (1) by visualizing the synthetic and real data points as clusters, one can visually verify if the synthetic samples are close to the real ones for each class and (2) check if the GAN model overfits the data, i.e., produces the exact same samples present in the training dataset. For all 24 categories, the PCA test shows that the clusters of synthetic samples and real

samples are close and overlap, thus validating the conditioning capabilities of the model. We show the PCA results for 2 categories in Figure 7. The PCA results for the remaining classes are presented in Appendix A.



(a) "unfriendly"



(b) "sure"

Figure 7: PCA results for two emotional states: (a) "unfriendly" and (b) "sure". The synthetic clusters of points are close to and overlap with the clusters of the real data points.

4.2 Subjective Validation

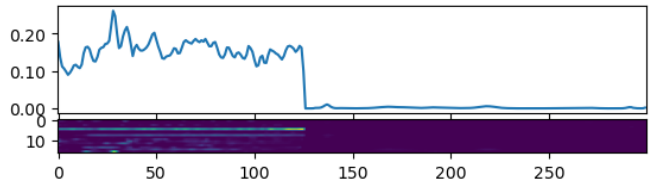
The protocol described in Section 3.5 was conducted with 11 participants. The mean age is 32.36 with a standard deviation of 8.80. 72% of the participants were male and 28% were female.

After collecting the results of the participants, we find that $P_{real} = 0.63$ and $P_{synthetic} = 0.52$. We note that out of the 55 attempts at classifying real samples by participants, 35 were classified as real, and out of the 55 attempts at classifying synthetic samples by participants, 29 were classified as real. Since the difference between the values is not very significant, one can conclude that the probability that a synthetic sample would be classified as real by human participants is not very different from the probability that a real sample would be classified as real by human participants. This suggests that our GAN model learned a probability distribution that approximates the real data probability distribution $p_{data}(x)$.

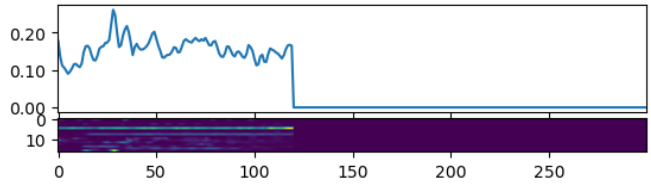
It must be noted that the subjective test was met with some difficulties, partly the lack of realism of the OpenFACS avatar, which introduces some subtle repeated movements to the lips. Additionally, many participants did not find the sequences very representative of the emotional states. This is due to the lack of information introduced by the absence of gaze and head movement information.

5 Limitations and Future Work

While this work provides a compelling case for using GANs to generate complex mental states for virtual agents, it is still lacking in certain aspects. Despite the realism of the generated synthetic samples, some of them exhibit some artifacts in areas where AU intensities are supposed to be null (see Figure 8a). A simple way of removing these artifacts is to use a running window that nullifies the intensities under a certain threshold. Using this method, the synthetic sample 8a could be corrected to 8b.



(a)



(b)

Figure 8: (a) shows a synthetic sample with artifact intensities in areas where it is supposed to be null. After applying a nullifying running window over the sequence, (a) was corrected to (b). Samples are scaled to the original $[0, 5]$ range.

Another limitation of this work is the subjective validation, which would have been more accurate if the facial expressions contained gaze and head movement information. Upon inspection of some emotional state videos, like "sad", we observed that much of the cues that indicate the state of sadness lie in the gaze and head movements. This was further validated by the participants who pointed out that some mental states were not very recognizable by the naked eye. Therefore, a good expansion of the current work would be to add gaze and head movement information for richer

and more realistic synthetic mental states. Additionally, a good future direction would be validating the realism of synthetic samples with a more extensive protocol with more participants, using more elaborate questionnaires like the Godspeed questionnaire [5], using a more realistic avatar for visualization.

6 Conclusion

In this work, we propose a conditional GAN model for generating realistic sequences of complex mental states. Despite the technical challenges that mainly lie in the scarcity of data, severe class imbalance and the variable-length nature of data samples, our model was successful in attaining good levels of realism, as confirmed by the proposed objective evaluation metric, the AUFID, the PCA test and the subjective validation. This work suggests that GANs show promise for animating virtual agents in a more realistic manner.

References

[1] Petar S. Aleksic, Gerasimos Potamianos, and Aggelos K. Katsaggelos. 2005. 10.8 - Exploiting Visual Information in Automatic Speech Processing. In *Handbook of Image and Video Processing (Second Edition)* (second edition ed.), AL BOVIK (Ed.). Academic Press, Burlington, 1263–XXXIX. https://doi.org/10.1016/B978-012119792-6/50134-0

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein Generative Adversarial Networks. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 214–223. https://proceedings.mlr.press/v70/arjovsky17a.html

[3] Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit. *IEEE International Conference on Automatic Face and Gesture Recognition* (2018).

[4] Simon Baron-Cohen and Jessica Kingsley. 2007. Mind Reading: The Interactive Guide to Emotions. *J Can Acad Child Adolesc Psychiatry* (2007).

[5] Christoph Bartneck. 2023. *Godspeed Questionnaire Series: Translations and Usage*. Springer, Cham, 1–35. https://doi.org/10.1007/978-3-030-89738-3_24-1

[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. 2019. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *The Seventh International Conference on Learning Representations*. ICLR. arXiv:1809.11096 [cs.LG]

[7] Sen Chen, Zhilei Liu, Jiaxing Liu, Zhengxiang Yan, and Longbiao Wang. 2021. Talking Head Generation with Audio and Speech Related Facial Action Units. *ArXiv* abs/2110.09951 (2021). https://api.semanticscholar.org/CorpusID:239024632

[8] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. CVPR. arXiv:1711.09020 [cs.CV] https://arxiv.org/abs/1711.09020

[9] Matthieu Courgeon, Céline Clavel, and Jean-Claude Martin. 2014. Modeling Facial Signs of Appraisal During Interaction; Impact on Users' Perception and Behavior. In *international conference on Autonomous agents and multi-agent systems (Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems)*. Paris, France, 765–772. https://hal.science/hal-01144001

[10] Vittorio Cuculo and Alessandro D'Amelio. 2019. OpenFACS: An Open Source FACS-Based 3D Face Animation System. In *Image and Graphics*, Yao Zhao, Nick Barnes, Baoquan Chen, Rüdiger Westermann, Xiangwei Kong, and Chunyu Lin (Eds.). Springer International Publishing, Cham, 232–242.

[11] Alice Delbosc, Magalie Ochs, and Stephane Ayache. 2022. Automatic facial expressions, gaze direction and head movements generation of a virtual agent. In *Companion Publication of the 2022 International Conference on Multimodal Interaction* (Bengaluru, India) *(ICMI '22 Companion)*. Association for Computing Machinery, New York, NY, USA, 79–88. https://doi.org/10.1145/3536220.3558806

[12] Alice Delbosc, Magalie Ochs, Nicolas Sabouret, Brian Ravenet, and Stephane Ayache. 2023. Towards the generation of synchronized and believable non-verbal facial behaviors of a talking virtual agent. In *Companion Publication of the 25th International Conference on Multimodal Interaction* (<conf-loc>, <city>Paris</city>, <country>France</country>, </conf-loc>) *(ICMI '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 228–237. https://doi.org/10.1145/3610661.3616547

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. CVPR, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[14] P. Ekman and W. Friesen. 1978. Facial Action Coding System: A Technique for the Measurement of Facial Movement. In *Consulting Psychologists Press*.

[15] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Networks. In *Proceedings of the twenty-eigth Annual Conference on Neural Information Processing Systems*. NIPS. arXiv:1406.2661 [stat.ML]

[16] Dan Hendrycks and Kevin Gimpel. 2023. Gaussian Error Linear Units (GELUs). *arXiv preprint arXiv:1606.08415* (2023). arXiv:1606.08415 [cs.LG]

[17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Proceedings of the the thirty-first Annual Conference on Neural Information Processing Systems (NIPS)*. NIPS, 6629 – 6640. arXiv:1706.08500 [cs.LG]

[18] Paolo Domenico Lambiase, Alessandra Rossi, and Silvia Rossi. 2023. A Two-Tier GAN Architecture for Conditioned Expressions Synthesis on Categorical Emotions. *International Journal of Social Robotics* (03 2023), 1–17. https://doi.org/10.1007/s12369-023-00973-7

[19] Jun Ling, Han Xue, Li Song, Shuhui Yang, Rong Xie, and Xiao Gu. 2020. Toward Fine-Grained Facial Expression Manipulation. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer International Publishing, Cham, 37–53.

[20] Zhilei Liu, Diyi Liu, and Yunpeng Wu. 2019. *Region Based Adversarial Synthesis of Facial Action Units*. Springer International Publishing, 514–526. https://doi.org/10.1007/978-3-030-37734-2_42

[21] Zhilei Liu, Guoxian Song, Jianfei Cai, Tat-Jen Cham, and Juyong Zhang. 2019. Conditional adversarial synthesis of 3D facial action units. *Neurocomputing* 355 (2019), 200 – 208. https://doi.org/10.1016/j.neucom.2019.05.003 Cited by: 17; All Open Access, Green Open Access.

[22] Birgit Lugrin, Catherine Pelachaud, Elisabeth André, Ruth Aylett, Timothy Bickmore, Cynthia Breazeal, Joost Broekens, Kerstin Dautenhahn, Jonathan Gratch, Stefan Kopp, Jacqueline Nadel, Ana Paiva, and Agnieszka Wykowska. 2022. *Challenge Discussion on Socially Interactive Agents: Considerations on Social Interaction, Computational Architectures, Evaluation, and Ethics* (1 ed.). Association for Computing Machinery, New York, NY, USA, 561–626. https://doi.org/10.1145/3563659.3563677

[23] Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, Zhen Wang, and Stephen Paul Smolley. 2017. Least Squares Generative Adversarial Networks. In *International Conference on Computer Vision (ICCV)*. arXiv:1611.04076 [cs.CV]

[24] A. Miolla, M. Cardaioli, and C. Scarpazza. 2023. Padova Emotional Dataset of Facial Expressions (PEDFE): A unique dataset of genuine and posed emotional facial expressions. *Behav Res Methods* (2023).

[25] Mehdi Mirza and Simon Osindero. 2014. Conditional Generative Adversarial Nets. arXiv:1411.1784 [cs.LG]

[26] Radoslaw Niewiadomski, Sylwia Hyniewska, and Catherine Pelachaud. 2009. Evaluation of multimodal sequential expressions of emotions in ECA. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. 1 – 7. https://doi.org/10.1109/ACII.2009.5349569

[27] Koichiro Niinuma, Itir Onal Ertugrul, Jeffrey F. Cohn, and László A. Jeni. 2022. Facial Expression Manipulation for Personalized Facial Action Estimation. *Frontiers in Signal Processing* 2 (2022). https://doi.org/10.3389/frsip.2022.861641

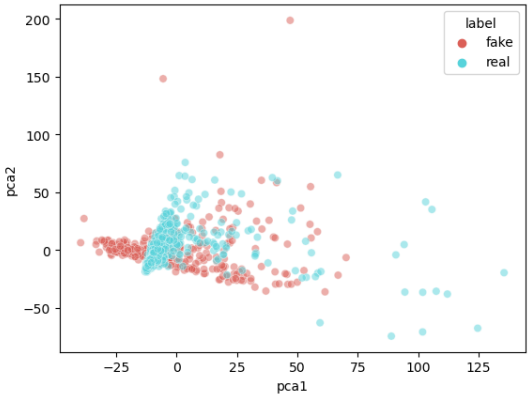
[28] Naima Otberdout, Mohamed Daoudi, Anis Kacem, Lahoucine Ballihi, and Stefano Berretti. 2019. Dynamic Facial Expression Generation on Hilbert Hypersphere With Conditional Wasserstein Generative Adversarial Nets. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44 (2019), 848–863. https://api.semanticscholar.org/CorpusID:198229775

[29] Catherine Pelachaud, Carlos Busso, and Dirk Heylen. 2021. *Multimodal Behavior Modeling for Socially Interactive Agents* (1 ed.). Association for Computing Machinery, New York, NY, USA, 259–310. https://doi.org/10.1145/3477322.3477331

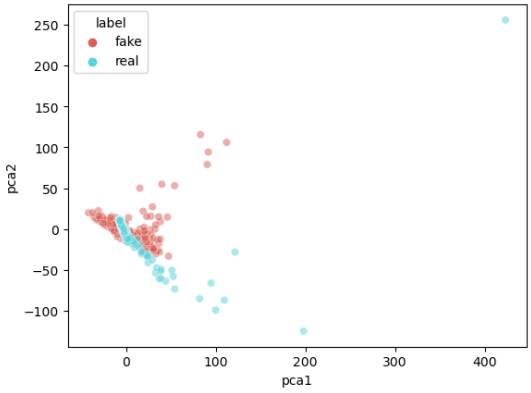
[30] Albert Pumarola, Antonio Agudo, Aleix M. Martínez, Alberto Sanfeliu, and Francesc Moreno-Noguer. 2018. GANimation: Anatomically-Aware Facial Animation from a Single Image. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part X (Lecture Notes in Computer Science, Vol. 11214)*. Springer, 835–851. https://doi.org/10.1007/978-3-030-01249-6_50

[31] Yong Zhao, Le Yang, Ercheng Pei, Meshia Cédric Oveneke, Mitchel Alioscha-Perez, Longfei Li, Dongmei Jiang, and Hichem Sahli. 2021. Action Unit Driven Facial Expression Synthesis from a Single Image with Patch Attentive GAN. *Computer Graphics Forum* (2021). https://doi.org/10.1111/cgf.14202

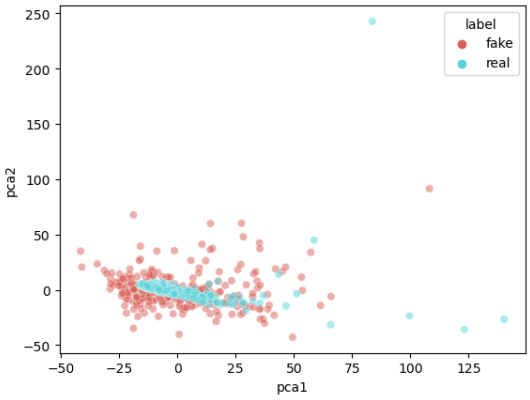
A PCA Results for the 24 Categories



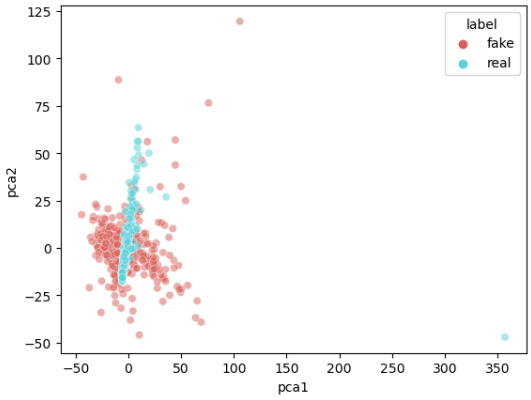
(a) "sad"



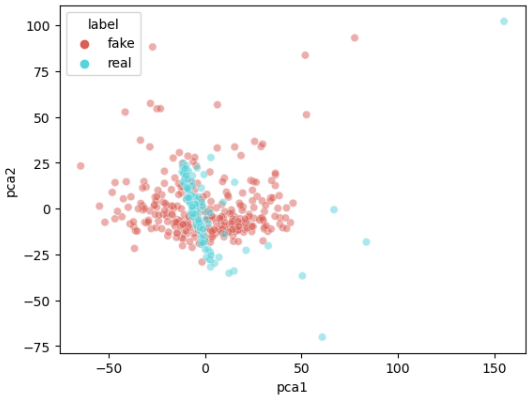
(b) "happy"



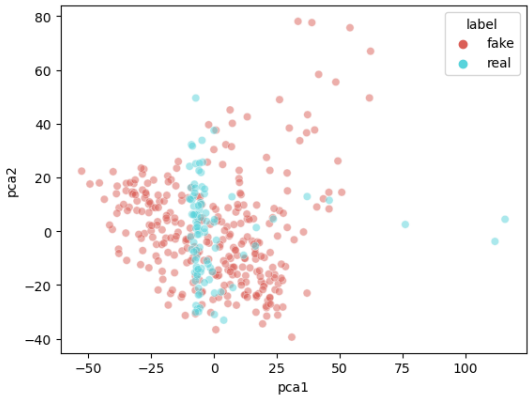
(c) "angry"



(d) "afraid"

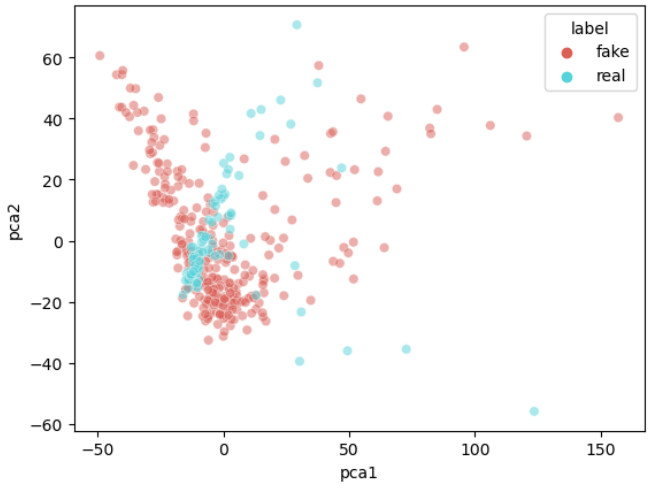


(e) "kind"

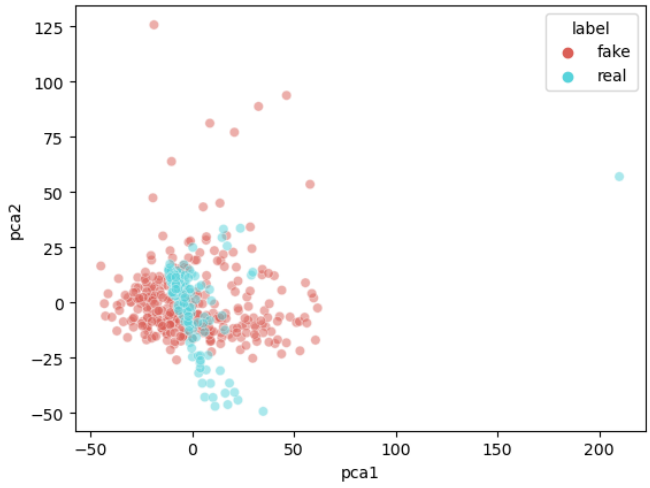


(f) "excited"

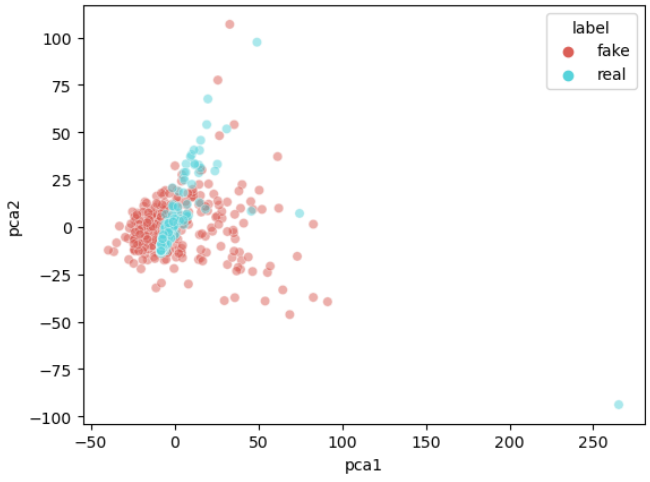
Figure 9: PCA results for all the 24 categories of emotional states.



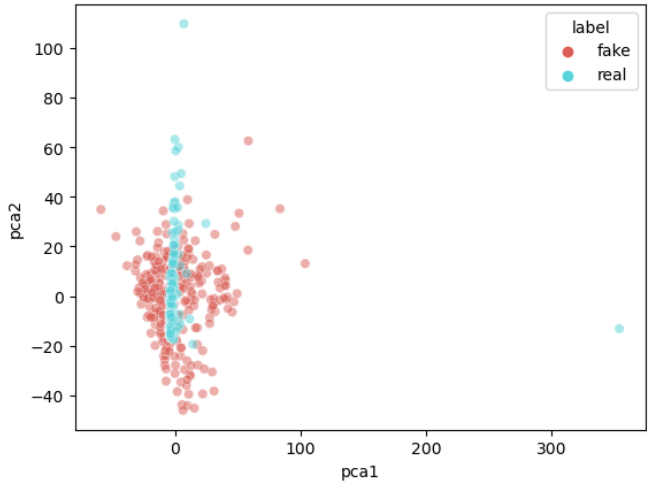
(g) "interesting"



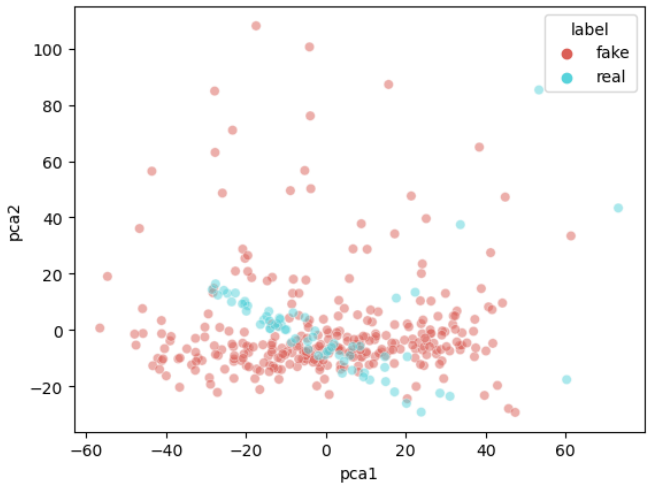
(h) "unsure"



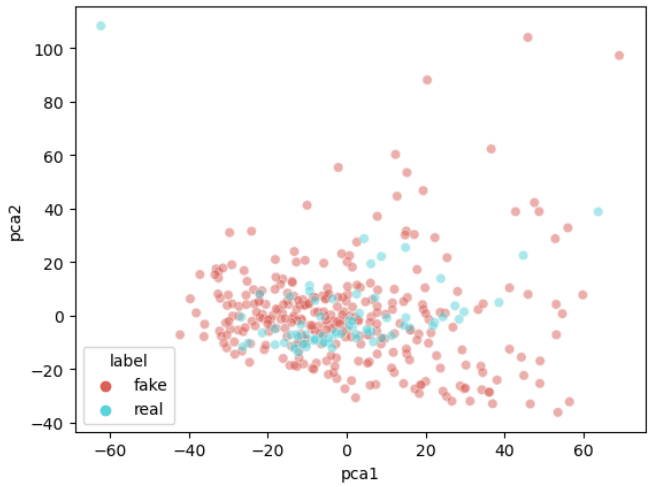
(i) "surprise"



(j) "hurt"

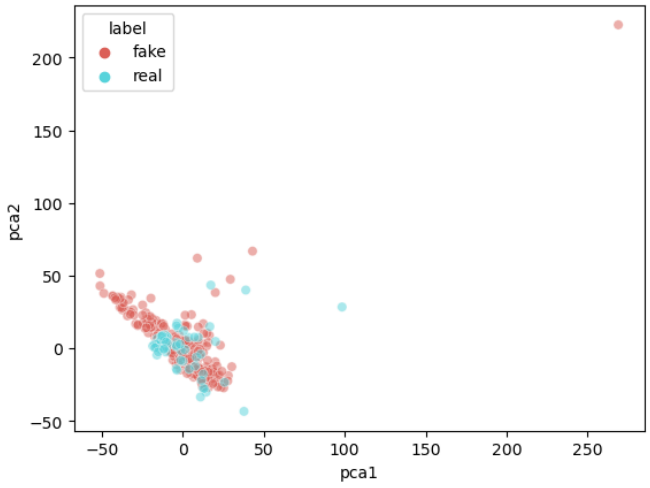


(k) "fond"

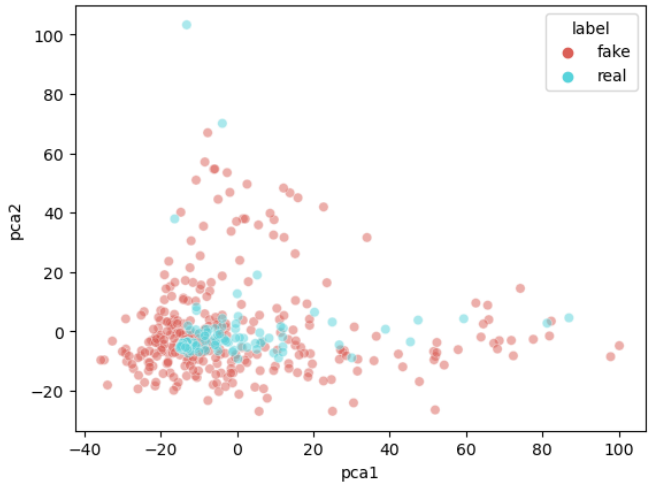


(l) "romantic"

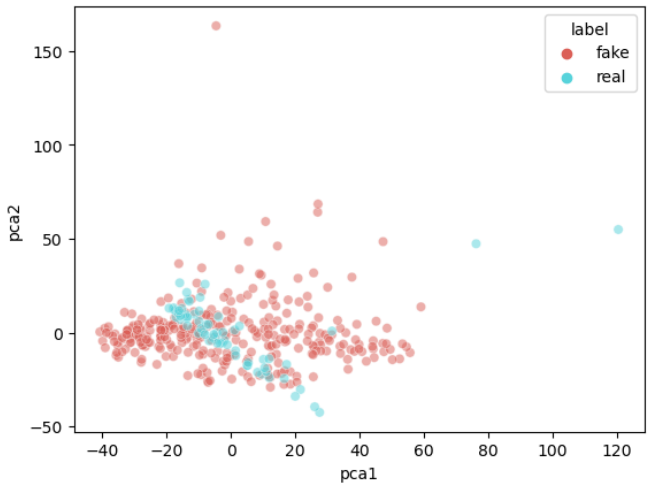
Figure 9: PCA results for all the 24 categories of emotional states.



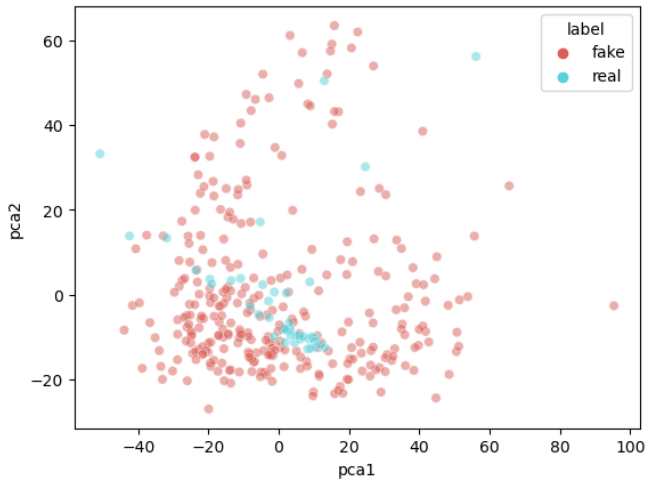
(m) "thinking"



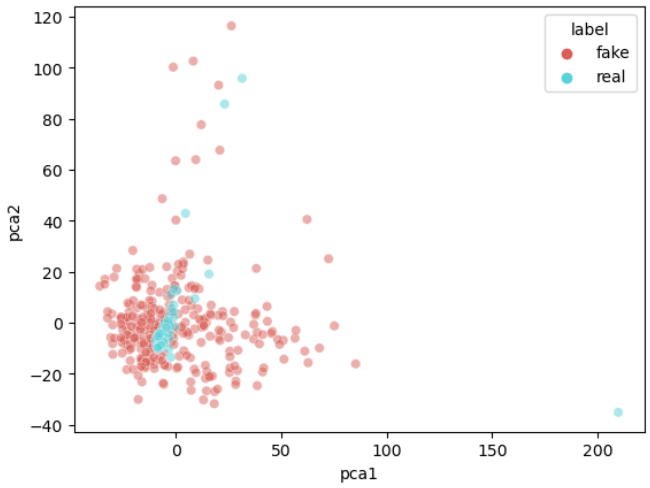
(n) "bored"



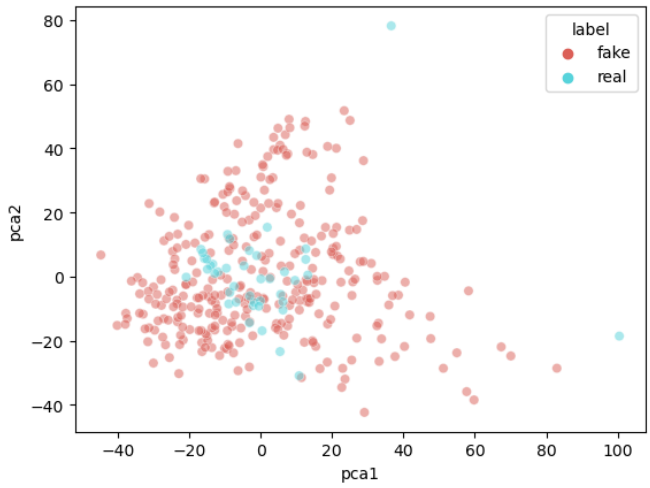
(o) "sneaky"



(p) "bothered"

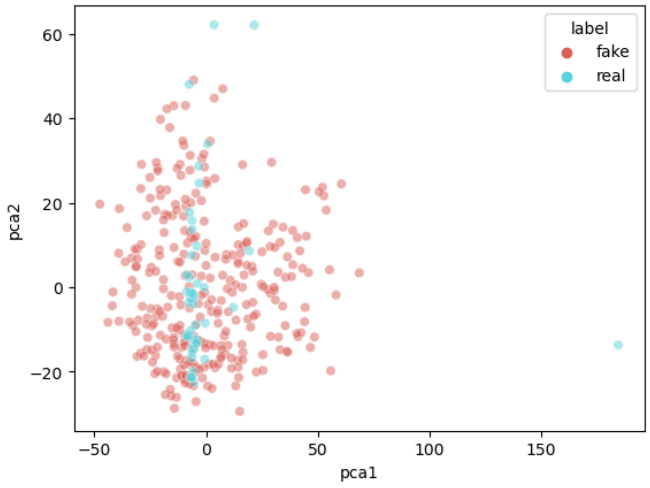


(q) "wanting"

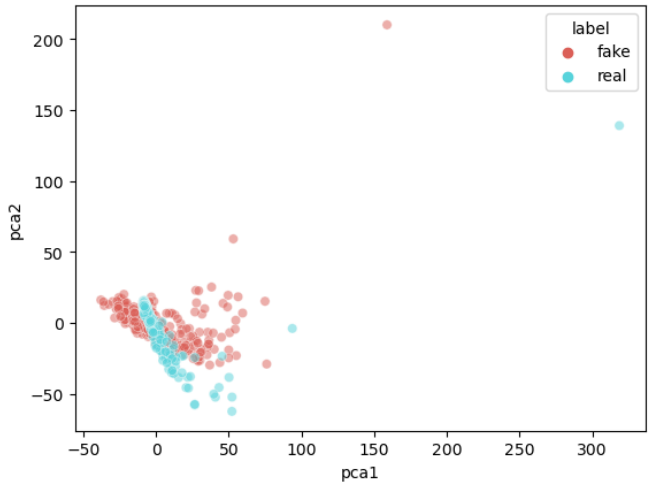


(r) "sorry"

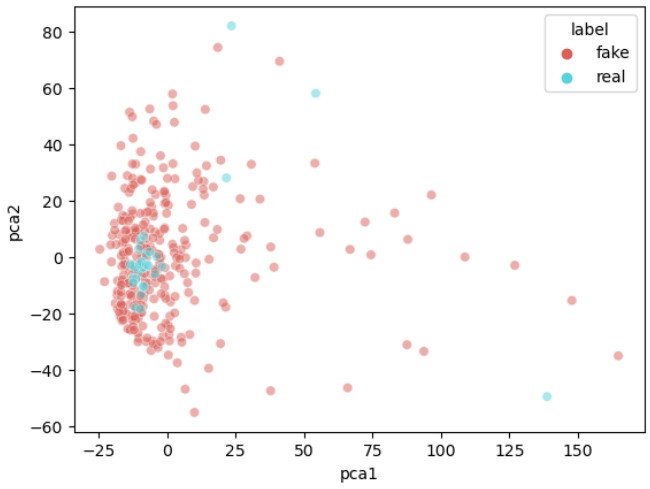
Figure 9: PCA results for all the 24 categories of emotional states.



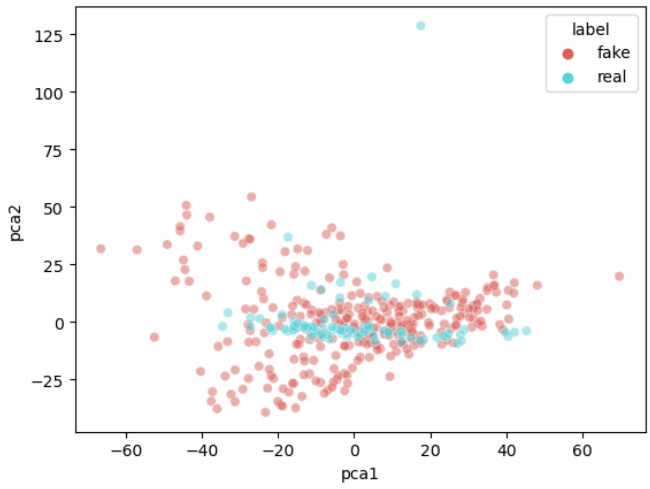
(s) "disbelieving"



(t) "disgusted"



(u) "touched"



(v) "liked"

Figure 9: PCA results for all the 24 categories of emotional states.