A Two-Tier GAN Architecture for Conditioned Expressions Synthesis on Categorical Emotions

Paolo Domenico Lambiase[1], Alessandra Rossi[1*] and Silvia Rossi[1]

[1*]Department of Electrical Engineering and Information Technologies, University of Naples Federico II, via Claudio, 21, Naples, 80125, Italy.

*Corresponding author(s). E-mail(s): alessandra.rossi@unina.it;
Contributing authors: plambiase94@gmail.com; silvia.rossi@unina.it;

Abstract

Emotions are an effective communication mode during human-human and human-robot interactions. However, while humans can easily understand other people's emotions, and they are able to show emotions with natural facial expressions, robot-simulated emotions still represent an open challenge also due to a lack of naturalness and variety of possible expressions. In this direction, we present a two-tier Generative Adversarial Networks (GAN) architecture that generates facial expressions starting from categorical emotions (e.g. joy, sadness, etc.) to obtain a variety of synthesised expressions for each emotion. The proposed approach combines the key features of Conditional Generative Adversarial Networks (CGAN) and GANimation, overcoming their limits by allowing fine modelling of facial expressions, and generating a wide range of expressions for each class (i.e., discrete emotion). The architecture is composed of two modules for generating a synthetic Action Units (AU, i.e., a coding mechanism representing facial muscles and their activation) vector conditioned on a given emotion, and for applying an AU vector to a given image. The overall model is capable of modifying an image of a human face by modelling the facial expression to show a specific discrete emotion. Qualitative and quantitative measurements have been performed to evaluate the ability of the network to generate a variety of expressions that are consistent with the conditioned emotion. Moreover, we also collected people's responses about the quality and the legibility of the produced expressions by showing them applied to images and a social robot.

Keywords: Generative Adversarial Networks, Conditional Emotion Expression, Action Units, Affective Computing, Social Robots

1 Introduction

Recent studies agree that robots designed to interact with adults and children according to social conventions tend to form stronger bonds with humans, in terms of trust, acceptance of collaborative tasks, and, generically, the success of the interaction [39, 42]. To facilitate such social interaction, robots should be able to execute efficiently physical and cognitive tasks (collaboratively or individually) and to look and behave as realistic, communicative, and effective [17]. A particularly important feature for facilitating Human-Robot Interactions (HRI) is the robot's ability to recognise, understand and reproduce human emotions. Emotion recognition and synthesis play a fundamental role in the effective communication between people and personal or service

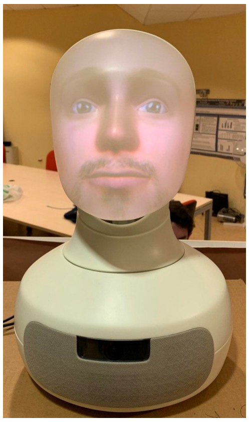


Fig. 1: Furhat robot by Furhat Robotics

robots [19]. Emotion modelling and computational approach for simulating affective behaviour in robots have been extensively investigated in the literature [8]. Still, many of these approaches rely on rule-based strategies to associate emotions with specific behavioural characteristics, such as kinesthetic, body posture and gestures, which make the behaviour of the robot predetermined given a specific emotion.

Moreover, the creation of realistic social robots, when they present a high level of anthropomorphism, still incurs the negative consequences of the uncanny valley [52]. While it is very common to have social robots with bodies that resemble human bodies (i.e., torso, arms, legs, stylised heads such as Softbank Robotics Pepper robot), the latest technologies allow robots to have replicated human faces with realistic features. An example of a robot with a human-like face is the Furhat robot whose facial features are projected in a plastic face-shaped case (see Figure 1). Such a robot's morphology presents the big challenge of generating the robot's natural affective facial expressions [16]. The generation of different and realistic facial expressions that can be appropriate and responsive to the contextualised human-robot interaction can be particularly problematic, especially while researchers are trying to enable robots to represent a wide and diverse population, such as gender, ethnicity, and age.

To tackle this issue, the use of artificial intelligence algorithms designed for the generation of realistic high-resolution images without encountering overfitting of the data, such as the Generative Adversarial Networks (GANs) [20] is a viable approach. GANs have been used in a huge variety of tasks, specifically for generating high-quality images with different gender and age, such as StyleGAN [25] and ESRGAN [53]. Generative Adversarial Networks have also shown great results for facial expression synthesis based on high-level attributes such as categorical emotions.

GANs aim is to generate realistic images from those of the training set. Given a training set, this technique learns to generate new data with the same statistics as the training set instead of simply reproducing it. This makes GANs an interesting approach to generating emotional expressions that vary from time to time. Indeed, we obtained very promising results for generating facial movements for expressing emotions from a human speech in our previous study [5]. These approaches, however, are very difficult to train [20], and they can generate a discrete number of expressions using the content and granularity of the dataset [35]. Moreover, the most relevant approaches based on GANs for directly generating faces introduce high computational costs, or just transfer an emotion across faces of different images [22].

To address these problems, we present a novel two-level GAN conditioning model that allows the generation of facial expressions based on discrete emotions (e.g., sadness, surprise, happiness) to have a high variety of different synthetic facial expressions for each emotion. The proposed approach allows encoding categorical emotions on a face decoupling the generation of low-level emotions from the generation of the whole face.

Our model is the result of the combination of a modified GANimation model [35] with a conditional GAN based on Action Units (AU). In detail, we realised a Conditional GAN (CGAN), called AU Generator, which takes in input a discrete emotion and generates a great variability of facial expressions to express such emotion. The ability to model various expressions for each emotion makes our architecture a useful tool for data augmentation by providing a solution to the problem of class imbalance of emotion-labelled human faces' dataset. In this work, we, therefore, refer to our architecture with the abbreviation AUGM (i.e., augmented) to highlight such capability. Once the AUs, conditioned on a discrete emotion, are generated, these can be applied to a robot face, like Furhat's one, or the GANimation can be used to apply them to a face for synthesising facial expressions in new images. In particular, we decide to modify the GANimation to hide the generation of AU vectors taking as input the discrete emotions as high-level attributes.

The replicability of the results presented in this study was guaranteed using three public datasets. We used the Facial Expression Research Group 2D Database (FERG DB) [2] composed of 2D images of stylised characters with annotated facial expressions, the AffectNet DB [31] that is the largest database of facial expressions, valence, and arousal with a million labelled images, and the CelebFaces Attributes Dataset (CelebA DB) [27] containing more than 200K celebrity images with 40 face attribute annotations.

The objective evaluation of a GAN's performance remains an open problem, and for this reason, we proposed several measures, using both automatic and human evaluations, to test the model's accuracy. The variety of images generated with the method presented in this work is used to enable a social robot to express various and more complex facial emotions. In particular, we used the synthesised AUs to easily and quickly generate a new set of emotion captures to be projected as facial features of the Furhat robot (also known as gestures)[1], and we conducted a preliminary investigation to compare people's ability to recognise the generated emotions as expressed by a human face and the robot. We aim as our next step to further investigate the effects of affective robotics in long-term human-robot interactions.

2 Related Works

In this section, we provide an overview of socio-affective modes used by robots to facilitate communication and interaction with humans, with a particular focus on the state-of-art techniques used for generating affective facial expressions, such as the models based on Generative Adversarial Networks (GANs).

2.1 Robot's Affective Competence for Social HRI

Emotional intelligence allows robots to create intuitive and natural interactions by granting them the ability to understand and use emotions for enhancing communication [47]. Humans are able to communicate several emotions by visibly varying their facial expressions, body and head movements, gestures and voice tone and pitch [44]. Therefore, affective computing has been applied in HRI by considering the same social signals. In the context of human-robot interaction, the first difficulty in generating robots' emotions is connected to the definition and formalisation of the emotions. Several models for describing emotions exist, and they are mainly varying between the categorical and dimensional models. Categorical models consist of discrete emotions associated with labels (e.g., sadness, happiness), while dimensional models consist of continuous values describing the emotions' features (e.g., arousal and valence). Spezialetti et al. [47] highlighted that there is not a clear agreement on which of these two models allows a better representation of human emotions. However, it is more difficult to identify emotions' features (i.e., dimensional model) compared to single emotion (i.e., categorical model). Moreover, most of the datasets available in the literature, which are essential tools for the recognition and generation of emotions, contain discrete emotions.

Several approaches [32, 40, 41, 45, 51] have been proposed in the literature, including classical machine learning or deep learning, that allow robots to communicate affective expression through body and head (e.g., open or close pose), gestures and movements (e.g., slow or fast). However, in artificial agents, a fundamental role is played by the possibility of expressing emotion through facial expressions [24]. According to a recent survey [38], there has been a gradual increase in research studies that aimed to recognise and reproduce emotions through facial expressions. For example, Xie and Hu [54] used a Deep-based Convolutional Neural Network for the recognition of facial expressions. Faria et al. [16] presented a dynamic probabilistic classification framework trained on the dataset Karolinska Directed Emotional Faces (KDEF) [28] for the recognition of facial emotions. Other works for the

[1] The Generation 2 version of the Furhat robots allows transforming the recordings of a face using a toolkit into a robot's gesture. More info at https://docs.furhat.io/gesture_capture_tool/

recognition of facial emotions used Support Vector Machine (SVM) [7], Deep Belief Network (DBN) [49], or Random Forests (RF) [55]. However, while few algorithms achieved good or high (up to 90%) accuracy also for the recognition of facial emotions in real-time and on dynamic input during HRI, the generation of facial emotions has been carried either hand-coded [4] or using Reinforcement Learning (RL) [11], GAN and GAN-based [12, 26] architectures.

2.2 GAN-based Architectures

The advance of Generative Adversarial Networks (GAN) obtained incredible results for tasks such as facial expression synthesis. In [48], Tang et al. presented EC-GAN, a conditional network that reproduces a given emotion by concatenating an emotional attribute to the vector representation of the image at the convolution layer level of the generator. The conditional attribute is represented in a categorical way. Another key point of this architecture is the use of a face mask loss that forces the generator to focus only on the image's region where the human face lies, preserving the background. Song et al. [46] proposed a GAN-based method, called Geometry-Guided Generative Adversarial Network (G2-GAN), that uses fiducial points for synthesising facial expressions. Geng et al. [18] combined the 3D Morphable Model (3DMM) and deep generative techniques generating a set of each expression for each real human face. However, these studies are limited to single frames or 2D representations. Moreover, these architectures present several limitations, such as instability of the training. Indeed, these approaches can generate a discrete number of expressions using the content and granularity of the dataset [35]. The most relevant approaches based on GANs for directly generating faces introduce high computational costs, or just transfer a defined emotional expression across faces of different images [22].

The most successful architectures in facial expression synthesis tasks that extend GAN's limitations are conditional generative adversarial networks (CGAN) [30], such as Star-GAN [10] and GANimation [35]. Star-GAN conditions GAN's generation process to a given attribute. The authors denoted the term attribute as a meaningful feature inherent in an image of a human face such as hair colour, gender, age or facial expression. A key feature of Star-GAN is a multi-domain image-to-image translation in a unified GAN. In other words, a single model can be trained on multiple datasets with different labels, learning to transfer feature attributes among themselves. Star-GAN uses a mask vector to achieve this result: an additional vector is added to the vector label. This vector is used to enable or disable dynamically the labels depending on the dataset the model is training on. Despite the performance in terms of photo-realism, Star-GAN can operate with discrete emotions but cannot ensure high variability in synthesised expressions by being able to change only a part of the face.

On the contrary, the approach using GANimation builds a GAN architecture that uses a conditioning scheme based on continuous facial movements represented by Action Units (AU). Given an emotional expression, encoded in terms of AUs, GANimation is able to efficiently synthesise the AUs on the input image. This architecture eliminates the problem of model collapse (since the class to which the generated data belong depends on the AU conditioning). Other benefits derived from the usage of numerical values to describe an expression (i.e., the AU) and from the use of a mechanism to focus the Generator only on the image's region involved in expression synthesis. Other regions such as hair, ears, glasses etc. are not modified by the net. For this purpose, GANimation uses an attention mask mechanism.

The objective of our work is to exploit the key point of this architecture and improve it by adding discrete conditioning with the automatic generation of AUs.

3 Approach

In this work, we distinguish expression and emotions, where an expression is meant to be the result of one or more motions or positions of facial muscles, while emotions are treated as a discretisation of facial expressions in the primary emotions (surprise, fear, happiness, anger etc.) [13]. The primary emotions taken into account in this work are based on Ekman's six basic emotion [13]. However, we introduced a "neutral" emotion instead of "disgust" due to the unavailability of sufficient data labelled with this emotion in the considered datasets. Hence, the discrete emotions considered

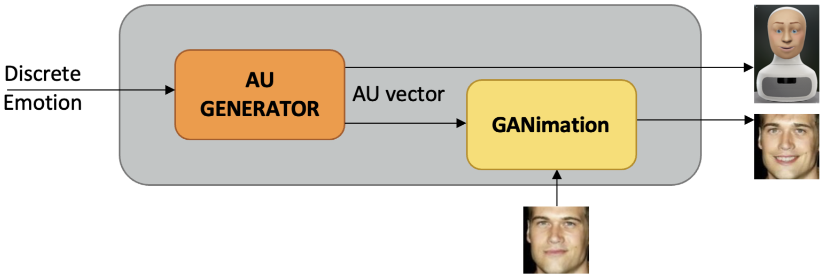


Fig. 2: The GAN architecture proposed in this work. The AU generator module architecture generates AU vectors that include the characteristics of an expression of the desired emotion (i.e., activation and facial muscle intensity) by taking as input a discrete emotion. The generated AU can be either directly applied on the robot's face or the GANimation module can be used to generate the facial expression by applying the characteristics expressed by the AU vectors on an image.

are: anger, neutral, fear, happiness, sadness, and surprise.

Facial expressions can be encoded by translating an expression into a set of values called Action Units (AU) using Facial Action Coding System (FACS) [14, 15]. The AUs are numerical vectors describing the fundamental facial muscles for each facial expression, whereas values greater than or equal to zero define respectively a contraction or a relaxation of one or more muscles of a specific face region. Since the determination of these values is not open to interpretation, they can be used for any higher-order decision-making process including basic emotions' classification. The underlying idea of our architecture is that variations of AUs correspond to the same emotion, and, therefore, the facial expressions' variability of emotion is linked to the generation of the different AUs describing such emotion.

Here, we proposed a two-module GAN architecture (see Figure 2), called AUGM. Its first module (AU Generator) is responsible for generating a facial emotion (in terms of AUs) starting from a discrete emotion (i.e., happiness, sadness, surprise, etc.), and the second module is a GANimation architecture to synthesise such emotion on a given input face. The first module is a Conditional Generative Adversarial Network (CGAN) used to generate AU vectors - which encode a face expression - for each class of emotions. The GANimation module creates a modified image that shows an expression by an AU vector and an image.

Hence, the input of the two-layer GAN architecture will be a given face and a discrete emotion resulting in applying this emotion to the face with a high variability due to the generation of different AUs corresponding to the discrete emotion. In our experimentation, the network takes in input images of size 128x128.

3.1 AU Generator Module

We used a CGAN for generating distinct AU vectors representing a certain class (i.e., Discrete Emotion). The conditional GAN learns a mapping between a latent space (noise) and the dataset data, and it guarantees the uniqueness of the vectors by the conditioning label representing a discrete emotion and a random noise vector, which is different at each execution.

The CGAN is composed of a Discriminator model and a Generator model, where the generator is used to synthesise fake data to resemble real data, and the discriminator is used to distinguish real and fake data. Both networks are deep networks composed of Fully Connected layers (FC).

3.1.1 Features Space

The network operates on vectors each one composed of 17 AUs. Such vectors have been retrieved from human face images labelled with emotions by using the OpenFace library. In detail, the information, including facial landmark, features, and AU detection, are retrieved using the OpenFace module called FaceLandmarkImg[2]. FaceLandmarkImg

[2] OpenFace module https://github.com/TadasBaltrusaitis/OpenFace

outputs a score of confidence indicating how confident is the tracker in the current landmark detection estimate. Hence, a filter is applied to use only AU vectors of samples where the confidence score is higher than a certain fixed value (0.9). Moreover, OpenFace generates 18 AUs for each expression, after an initial empirical evaluation of results, but we noted that AU9 was responsible for many artefacts on the final image generations, so AU9 will not be considered. The values of AU activations are in the range (0,..,5). The AU vector is embedded with the label of the emotions shown in the samples. The classes are composed of the six main basic emotions [13]: [neutral, happy, sad, surprise, fear, and anger]. The network uses the index of this array to determine the classes (e.g., 0 is neutral, 1 happy etc.).

3.1.2 AU Generator

The generator uses the noise vector z and the label y to synthesise the fake example $G(z, y) = (x^*|y)$ (x^* given the label y). The goal is then to generate a fake example that is as close as possible to a real example belonging to the y label.

The generator is composed of six fully connected layers. It takes a noise vector embedded with a label vector as input, and each hidden layer has a number of neurons that is double the layer that precedes it. The output layer is composed of 17 neurons, one for each AU, and it returns a synthetic AU vector. The activation function of the input layer and the hidden layers is a 0.2 Leaky ReLU, while the output layer is the classical ReLU, to force the values of output to be not negative. A Batch-Normalisation with momentum $\alpha = 0.8$ is applied to the input and each hidden layer. The momentum reduces the noise in the gradient update term, which helps a faster convergence towards the optimal (or near-optimal) value. The Batch-Normalisation can reduce the coupling of the layers' parameters, thereby stabilising the input of the layer, and consequently increasing the speed of convergence. It might be difficult to track the mean (π_B) and variance (σ_B^2) of a batch during normalisation, and it is needed to update the batch mean and variance with an exponentially weighted "moving average". In particular, during the training process, the moving-mean (π_{mov}) and moving variance (σ_{mov}) are calculated with Equations 1.

$$\begin{aligned} \pi_{mov} &= \alpha \pi_{mov} + (1-\alpha)\pi_B \\ \sigma_{mov} &= \alpha \sigma_{mov}^2 + (1-\alpha)\sigma_B^2 \end{aligned} \tag{1}$$

The generator learns the mapping between a vector $z + y$, where z is the noise vector and y the label vector, and the vector AU. The network generates an AU vector that belongs to the class indicated by y. The models for the Generator networks used for implementing our CGAN are shown in Table 1.

Table 1: Generator's model of the AU Generator.

Layer	Input → Output Shape	Layer Information
Input layer	dim_{latent} + $n_{emotions}$ → 128	FC - Leaky ReLU (0.2) + Batch-Norm(momentum = 0.8)
Hidden layer	128 → 256	FC - Leaky ReLU (0.2) + Batch-Norm(momentum = 0.8)
Hidden layer	256 → 512	FC - Leaky ReLU (0.2) + Batch-Norm(momentum = 0.8)
Hidden layer	512 → 1024	FC - Leaky ReLU (0.2) + Batch-Norm(momentum = 0.8)
Hidden layer	1024 → 2048	FC - Leaky ReLU (0.2) + Batch-Norm(momentum = 0.8)
Hidden layer	2048 → 4096	FC - Leaky ReLU (0.2) + Batch-Norm(momentum = 0.8)
Output layer	4096 → n_{AU}	FC + ReLU

3.1.3 AU Discriminator

The input of the Discriminator is composed of the AU and the label vectors (x, y) from real examples and fake examples along with the label $(x^*|y, y)$ produced by the Generator. The goal of the Discriminator is to learn to reject all fake examples and all examples that do not match the given label and to accept all pairs of real examples that match. Therefore, the discriminator takes an AU x vector and a y label in input and returns the likelihood that the input is a real, matching pair. The activation function of every layer, except the output layer, is a 0.2-Leaky ReLU. The hidden layers use a 0.4-dropout. The α-Leaky ReLU is an activation function derived from the ReLU function. The α-Leaky ReLU is defined in Equation 2.

$$\alpha \text{ Leaky ReLU} = \begin{cases} x, & \text{if } x > 0 \\ \alpha x, & \text{otherwise} \end{cases} \tag{2}$$

The Leaky ReLU function has been preferred to the classic ReLU function because it is not rare for the ReLU function to suffer from the "Dying ReLU" phenomenon in a deep neural network. A dead neuron is a neuron with a ReLU activation function that has been set to 0 and never changed its value. Such neurons are considered useless because they do not give any contribution during the training phase. Contrarily, Leaky ReLu has a small slope (α) for negative values, instead of zero, and it produces faster training. The ρ-dropout is used to choose a random set of neurons that are ignored during the training process. At each training iteration, every node is dropped out with a probability ρ or kept with a probability of $1 - \rho$, forcing neurons within the same layer to take on more or less the responsibility for the inputs in a probabilistic way. In a fully connected layer, the neurons typically develop a co-dependence between them during training. This co-dependence, which is an essential part of the training process, should not prevent individual neurons from "relying" completely on each other, which can lead to overfitting the training dataset. The model for the Discriminator network used for implementing our CGAN is shown in Table 2.

Table 2: Discriminator model of the AU Generator.

Layer	Input → Output Shape	Layer Information
Input layer	n_{AU} + $n_{emotions}$ → 512	FC - Leaky ReLU (0.2)
Hidden layer	512 → 512	FC - Leaky ReLU (0.2) + dropout (0.4)
Hidden layer	512 → 512	FC - Leaky ReLU (0.2) + dropout (0.4)
Hidden layer	512 → 512	FC - Leaky ReLU (0.2) + dropout (0.4)
Hidden layer	512 → 512	FC - Leaky ReLU (0.2) + dropout (0.4)
Hidden layer	512 → 512	FC - Leaky ReLU (0.2) + dropout(0.4)
Output layer	512 → 1	FC

3.1.4 Training Process

The training process of our CGAN network has been organised in two steps, one for the Discriminator model and one for the Generator model.

The Generator's training step takes in input a batch of noise vectors and a batch of random labels, and then the discriminator is asked to predict the generator loss on the generator's output, which will be back-propagated to tune the generator's weights.

The Discriminator training step predicts a first time on a batch of samples coming from the dataset and a second one on a batch of synthetic samples. The losses of the two outputs are calculated and summed to constitute a final loss that is back-propagated to update the Discriminator weights. The loss function of the net is set to mean squared error according to Xudong Mao et al. [29]. This approach is based on the observation that the usage of binary cross-entropy does not guarantee the generation of samples that look real. This could happen because the binary cross-entropy leads to very small or vanishing gradients, and the model uses fake samples on the correct side of the decision boundary even if they are still far from the real data. The Discriminator uses the loss to minimise the "sum squared error" between predicted and expected values for real and fake samples. In contrast, the Generator minimises the "sum squared" difference between predicted and expected values for generated images. The Discriminator and Generator use the equations shown in Equation 3.

$$\begin{gathered} Discriminator : min(D(x) - 1)^2 + (D(G(z)))^2 \\ Generator : min(D(G(z)) - 1)^2 \end{gathered} \tag{3}$$

In this approach, the Discriminator predicts the class labels of 0 and 1 for fake and real images respectively, minimising the least-squares, called mean squared error. The mean square error is calculated according to Equation 4

$$MSELoss : \sum (y_{pred} - y_{true})^2 \tag{4}$$

where y_{pred} is the predicted class (Fake/Real) and y_{true} is the ground truth.

The loss is used during the conditioning by maintaining the class labels 1 for real samples that belong to the conditioning class in the input and 0 for fake samples or samples that do not belong to the conditioning class. This allows the model to learn how to generate samples that are both realistic and belong to the input class by using the conditioning class (i.e., emotion).

We then use the Adam optimiser to compute an efficient stochastic optimisation that only requires the first and second moments of the gradients with a small memory requirement.

3.2 GANimation Module

One of the key features of GANimation is to focus only on specific regions or characteristics (such for example colour) of an image that are related to the synthesis of the expression leaving the others unaltered. This is done by the use of a so-called attention mask within the generator. Hence, the GANimation applies an attention and an activation mask on the image and the AU vector received in input from AU Generator Module. The final image produced is the result of the overlapping of the two masks.

The GANimation module is also composed of a Discriminator and a Generator model. The Generator model is composed of thirteen layers. The first three levels of down-sampling are implemented through convolution layers to extract features from the image. The next 5 levels are residual blocks, also convolutional, that transport the information from the previous layers to the next layers. This process allows a more stable training process by limiting the problem of gradient vanishing and consequently propagating larger gradients to the initial layers. Therefore, the use of bottleneck layers allows an alternative path for the gradient through back-propagation. The next 4 layers are up-sampling and they perform deconvolutions to reconstruct the final image. These 4 layers are composed of the first two layers in cascade and the last 2 layers in parallel. The last two layers are the output layers. The first of these two layers (Up-sampling ← with the mask) returns the colour mask, that describes the face modified according to the AU vector. The second output layer (Up-sampling → att mask) returns the attention mask, i.e., the filter that describes the areas of the initial image to be modified with the colour mask.

The Discriminator model is composed of eight layers. The first layers, an input layer and 5 hidden layers are convolutional. The hidden layers are followed by two parallel output layers. The first final layer (Output Layer (AU reg)) performs a regression estimating the AU of the input image. The second final layer (Output Layer (Classification)) performs the task of the critic, returning a truthfulness score to the initial image, according to the PatchGAN procedure. The procedure consists of dividing the image is 64 patches and predicting the probability that each patch is real or fake. The final score is, then, obtained by calculating the mean of every patch's probability.

We used a Wasserstein distance (WGAN-GP) with gradient penalty and instance normalisation as GAN loss function for both the generator and discriminator [3]. The use of WGAN-GP allows a more performing instance normalisation than batch normalisation due to the use of the gradient penalty, which must be independently imposed on different samples.

4 Datasets

We used three different datasets for testing our neural network which were selected according to the following characteristics: 1) they need to contain images of human faces; 2) they need to be annotated according to the discrete emotions of the human faces that have been used to train the network to synthesise data conditioned on the discrete emotion; and 3) they need to contain information about the AU values of the faces in the images. The three datasets selected are:

FERG DB

Aneja et al. created FERG B [2], a database composed of six stylised characters labelled by facial expressions (see Figure 3a). The stylised characters are 3 males and 3 females (Ray, Malcolm, Jules, Bonnie, Mery and Aia). The characters were modelled in 3D using the graphical software called MAYA and rendered in 2D to produce the final image. The images for each character are grouped according to seven main emotions: Anger, Disgust, Fear, Happiness, Neutral, Sadness and Surprise. The animator created the key poses for each emotion, and labelled them, to populate the database.

On average, the authors created 150 key poses (15-20 per emotion) for each character, and then they interpolated them for creating a generalised feature space among the characters. The final images produced are 55767 images (8000 per character).

AffectNet DB

AffectNet [31] contains about 1M of facial images collected from the Internet by querying three major search engines using 1250 emotion-related keywords in six different languages (see Figure 3b). About half of the images retrieved are manually annotated based on seven discrete facial expressions (categorical model) and the intensity of valence and excitement (dimensional model). The rest of the images are annotated automatically using the trained ResNet neural network on all samples of the training set with manual annotation with an average accuracy of 65%. AffectNet is by far the largest database of facial expressions, valence and excitement, available for automatic recognition of facial expressions in two different emotion patterns. As for the discrete emotions (categorical model), the images are annotated according to eleven categories of emotions as follows: 0: Neutral, 1: Happiness, 2: Sadness, 3: Surprise, 4: Fear, 5: Disgust, 6: Anger, 7: Contempt, 8: None, 9: Uncertain, 10: Faceless. In particular, the category None ('None of the eight emotions') cannot be assigned by the annotators to any of the other basic emotions. However, valence and arousal values can be assigned to these images. The faceless category is used to label images without a human face or with not clearly recognisable faces (e.g., distorted faces). The images annotated with Uncertain are those for which the annotators were unsure which emotions to assign.

CelebA DB

CelebFaces Attributes Dataset (CelebA) [27] is a large-scale face attributes dataset with more than 200K celebrity images, each with 40 attribute annotations (see Figure 3c). The images in this dataset cover large pose variations and background clutter. CelebA includes 10177 identities, 202599 face images, 5 landmark locations, and 40 binary attributes annotations per image. The dataset can be employed as the training and test sets for the following computer vision tasks: face attribute recognition, face detection, landmark localisation, and face editing and synthesis.

5 Experimental Results

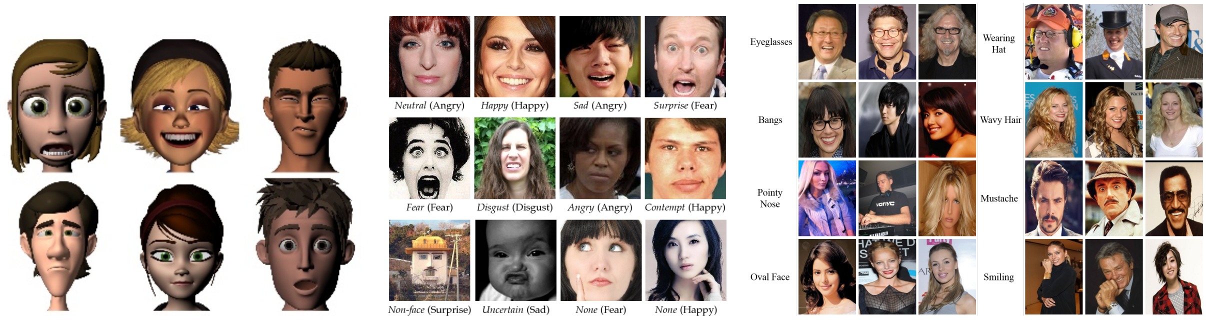
The test plan of the GAN's performances was carried out by first evaluating the AU Generator results, and then by evaluating the quality of the final images. A sample of synthesised emotions generated by our architecture, which evaluation is discussed in this section, is shown in Figure 4.

In particular, the test plan has been divided into the following steps:

1. AU Generator Module test:
   - We used the GAN-test approach to analyse the convergence of GAN training. Every n steps of training, a pre-trained AU-Emotion Classifier is launched on a batch of AU-generated vectors to test the accuracy of the GAN conditioning;
   - The average and standard deviation of the AU values of the original and generated data are calculated for testing the diversity and realism of the generated data. The values of both generated and original data are compared to inspect the ability of the GAN to replicate the key features of the original dataset.
2. AUGM test:
   - Test the quality of the overall images produced using Frechet Inception Distance (FID);
   - Test the quality of the conditioning using a pre-trained emotion classifier;
   - Test the quality of both images and conditioning with a web-based interview.
   - Test the recognition of the produced emotion expression once applied on the Furhat face with a web-based interview.

5.1 AU Generator Module Test

In this section, we discuss the metrics used for evaluating the performance of the AU Generator module. For this phase of testing, we extracted the information from the datasets related to the activation and intensity of facial muscles (AUs) and the emotion labels. The tests were performed using PyTorch on CUDA with an nVidia GeForce GTX 1080Ti. All the tested models were trained for 10000 epochs.



(a) *Samples of FERG characters.* **(b)** *Samples of queried images from the web and their annotated tags [2].* **(c)** *Samples from CelebA*[3].

Fig. 3: Samples of the datasets (a) FERG, (b) AffectNet and (c) CelebA used for this work.

5.1.1 Conditioning Test

We tested the convergence by training a classifier that predicts the emotions for a certain AU vector. The classifier is interrogated on a batch of synthetic AU vectors every 50 training epochs. We calculated the accuracy of the classifier after each interrogation and compared the accuracy of the classifier on the original test set and the accuracy of the classifier on the fake test set.

The multi-class accuracy used to evaluate the model is defined as follows: the classifier in question returns, for each class, the probability that the input data belongs to that class. To compute the accuracy, which is the ratio between the number of well-classified samples and the total number of elements of the test set, an element is defined to be well-classified when the out probability to be in the right class is above 50%.

The datasets were randomly divided into the training set, validation set and test set according to the 80-10-10 schema. That is, 80% of the data will be used as the Training set, and the remaining 20% divided equally between the validation set and the test set.

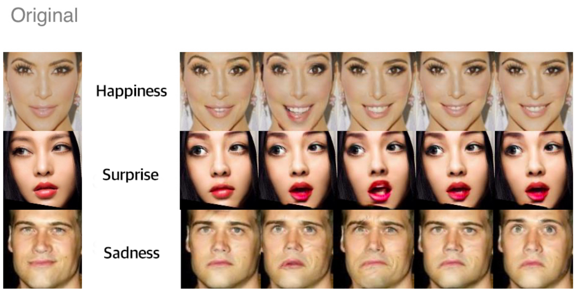


Fig. 4: Samples of images showing Happiness, Surprise and Sadness emotions generated by the proposed GAN-based architecture.

The structure of the network is composed of two hidden layers one input layer and one output layer (see Table 3). The activation function for the input layer and the hidden layers are Leaky ReLU. The output layer, instead, uses a softmax activation function. The hidden layers also use a 0.4 dropout. The loss function used to train the net is the categorical cross-entropy.

Table 3: AU Generator's Discriminator model.

Layer	Input → Output Shape	Layer Information
Input layer	$n_{AU} \rightarrow 512$	FC - Leaky ReLU (0.2)
Hidden layer	$512 \rightarrow 512$	FC - Leaky ReLU (0.2) + dropout (0.4)
Hidden layer	$512 \rightarrow 512$	FC - Leaky ReLU (0.2) + dropout(0.4)
Output layer	$512 \rightarrow 1$	FC + softmax

The classifier was initially trained on the AffectNet dataset obtaining an accuracy value of 87.7% on the test set. The AU Generator was then tested using the prediction of the trained classifier, but the results were not in line with what was desired. We observed little stability during the training, with an upper bound of 59% of accuracy,

[3] Picture's source: http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

which indicates that training with this dataset has a serious conditioning problem. We believe that this is due to the heterogeneity of the dataset and the images' quality. This needs to be considered for the next phase because it might introduce an error in the determination of AU vectors by the Openface module and make the association with the AU-label vector inconsistent.

We believe that it is preferable to train the model on a dataset whose images are clearly marked with emotions. The main reason lies in the fact that the AU Generator module's work is abstracted from the image itself and focuses solely on the encoding of the emotion represented in terms of AU vectors. For this reason, we tested the model on the FERG dataset, which contains only 6 subjects, it has a multitude of expressions for each emotion for each character. Moreover, the dataset is formed by avatars modelled in 3D and reported in 2D, and, therefore, given the cartoon nature of the subjects, marked with evident expressions. We trained the classifier on the FERG dataset obtaining an accuracy of 95.7%, which is higher than the accuracy of the classifier in the AffectNet dataset (87.7%) and resulting in a stronger and clearer AU vector-Emotion relationship. We observed that the model trained with the FERG dataset learned much faster than the model trained with AffectNet to generate data belonging to the desired class. Moreover, results showed that the level of accuracy is identical to that of the classifier on the original test set.

5.1.2 Realism and Variability Test

We then tested both the realism and the variability of the data produced. We calculated the means of the values both of the original dataset and a batch of 6000 generated vectors (1000 per emotion) for each emotion and AU. We calculated, in the same way, the standard deviations for each emotion and AU of both the original dataset and the batch of fake vectors. These metrics should not be confused with a simple evaluation of how much the network is replicating the dataset, because the network cannot simply do this job by its very nature. Here, we use the average values to estimate the goodness of the network in capturing features of the original dataset without reproducing it. For example, the network's learning ability is that a certain emotion is more likely to have a certain AU activity and intensity. Hence, differences between the averages and the standard deviations gave us a general metric of how much the network captured such key features of the original dataset and learned to generalise.

We also compared the differences in the average and standard deviation of the model trained on the FERG and AffectNet datasets. In Figure 5, we can observe that the model trained on the FERG dataset is generally better than the one trained on AffectNet, and that the network is able to learn better both to capture the key features of the original dataset (by observing the average) and to replicate its distribution in terms of data variability (by observing the standard deviation values). These are in line with the results on conditioning showing that a dataset with cartoon-like images, with respect to a wider and more realistic one, makes it easier for the network to learn AUs' values distributions for every single emotion.

We also ran a one-way ANOVA test to inspect the differences between the fake dataset and the original dataset (i.e., if they have the same stochastic distribution). The ANOVA is carried out for each basic emotion and AU, considering the average AU-scores for each emotion. The results are shown in Table 4. We did not observe any outliers, as assessed by the boxplot, and the data were normally distributed. The two metrics presented are F and p, where F represents the ratio between the "between" variance and the "among" variance, and p represents the probability of the null hypothesis is true. The result of the test suggests that there is insufficient evidence to reject the null hypothesis in the case of the FERG dataset, and then any observed difference in the model is likely due to statistical chance. The same does not hold for the AffectNet dataset where the null hypothesis can be rejected (with the exception of Happiness and Fear) and the observed difference is likely due to a difference in the generated model w.r.t. the original dataset. Indeed, the results are in line with our previous test validating that the model produced better accuracy on FERG than AffectNet datasets.

5.2 AUGM Module Test

We evaluated the GANimation module using three tests. We first used the Frechet Inception Distance to evaluate the ability of the models to work in

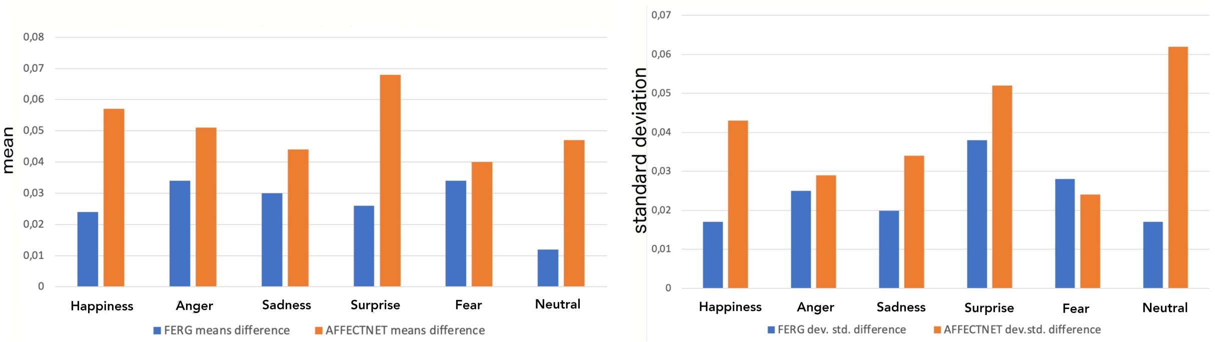


Fig. 5: Comparison of the differences of means of the AUs' values between original dataset and fake samples (figure on the left); standard deviation between original dataset and fake samples (figure on the right).

Table 4: ANOVA statistical F and p values of the differences between the FERG and AffectNet datasets with the original dataset.

	FERG dataset		AffectNet dataset	
Emotion	F	p	F	p
Happiness	7.47	0.16	37.53	0.10
Anger	13.48	0.22	248.58	0.03
Sadness	8.30	0.22	176.12	0.01
Surprise	41.51	0.17	24.12	0.01
Fear	16.51	0.14	18.34	0.13
Neutral	6.50	0.18	12.02	0.11

synergy and produce photo-realistic results. This distance assesses both the quality of the data in terms of photo-realism and the diversity of the generated data. Therefore, indirectly, it represents a further test of the number of expressions that the AUGM is able to synthesise. The second phase involves the creation of an image-based emotion classifier to evaluate the conditioning of the network, this time, unlike the test on the AUGM, in terms of final produced images.

We tested the model by training GANimation on two separate datasets, CelebA and AffectNet, both for 100 epochs, and we compared the results. We also compared the application of a vector composed of the available AUs (i.e., 17) and vectors composed of subsets (i.e., 7 and 12) of these to evaluate how much the total number of active AUs affects the final quality of the photo and the ability of the network to express the desired emotion. The subset of AUs used in the tests has been chosen empirically by selecting the subset leading to better (e.g., lower) score values. The tests were performed using PyTorch on CUDA with a nVidia GeForce GTX 1080Ti, and the training process of the AUGM takes 3-4 days for 100 epochs. Finally, we conducted a user study to evaluate our model in terms of recognition of emotion.

5.2.1 FID Scores

The FID score is a metric that evaluates the ability of a GAN to produce realistic and various samples. We used a sample of 13000 images from the original dataset and 7000 generated images since it is necessary to provide the classifier that will compute the distance with no less than 10000 samples to obtain meaningful results [23]. The tests were executed by training the model on CelebA and AffectNet separately and evaluating the trained model using samples from both datasets. Lower values for FID correspond to more similar real and generated samples as measured by the distance between their activation distributions, where 0 is ideally the best score, and $+\infty$ is the worst score.

We compared the values obtained with the proposed architecture to ones obtained from some general generative tasks, because similar architectures, such as GANimation or StarGAN, used humans to conduct a realism test [1]. Tables 5 and 6 show the FID values of our GANimation module trained on AffectNet, CelebA, and those produced from different models. The results of our module are lower than the other models, but the task here was not strictly generative. We do not intend, however, to directly compare the values, but to give an idea of the goodness of the values of this test. The results also show that training with the AffectNet dataset gives better FID results than training with the CelebA dataset. In particular, we can observe that the images modified by taking into account a lower number of AU (and thus muscles' activation) are less prone to distortion (7 AUs). However, this also leads to a misinterpretation of the expression represented.

Table 5: Comparison of the FID values produced from similar GANimation models. Lower values for FID indicate more similar real and generated samples.

Approach	FID
Gated PixelCNN [33]	65.9
DCGAN [36]	37.1
Coulomb GAN [50]	27.3
TTUR [23]	24.8
MoLM [37]	18.9

Table 6: FID values produced from the GANimation module trained on AffectNet and CelebA datasets. Lower values for FID indicate more similar real and generated samples.

GANimation module trained on AffectNet			
Datasets	7 AUs	12 AUs	17 AUs
CelebA	8.70	10.12	11.03
AffectNet	10.33	10.98	11.27
GANimation module trained on CelebA			
Datasets	7 AUs	12 AUs	17 AU
Celeb A	13.24	14.43	15.57
AffectNet	18.87	19.87	20.04

We can also notice the AffectNet-trained model achieved better results when the test images come from CelebA than AffectNet. We believe that CelebA can count on sharper images with generally better visual quality than AffectNet. AffectNet has many more images with different facial poses and more expressions. This makes the models trained on AffectNet better than those trained on CelebA in terms of resulting visual quality. Whereas the best results are therefore achieved with the combination "trained on AffectNet - Images from CelebA" because the model has benefited from the training on AffectNet and the test images are high-quality CelebA images. An example of the differences between AffectNet and CelebA training is shown in Figure 6.

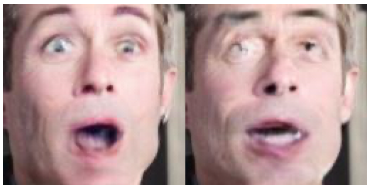


Fig. 6: Difference between AffectNet and CelebA training. The same AU vector corresponding to the Surprise emotion is applied to the same image. The image on the left is produced by the model trained on AffectNet, and the image on the right is generated by the model trained on CelebA. The image on the left is more defined and less blurred than the other, and thus the target expression is more clearly recognisable.

5.3 Emotion Classification Test

As a second case, we decided to test the accuracy of our model by evaluating the conditioned generation of AUs, once applied to a final image, in terms of automatic emotion recognition. To do so, we chose to train a classifier, which takes an image that represents a human face as input and predicts the relative expressed emotion. The classifier used was a ResNet architecture [21] a deep network composed of 152 layers and residual blocks, and we trained it on a subset of the AffectNet dataset. In particular, we selected images for which LandmarkImg of OpenFace provided a confidence score in landmark detection estimation greater than 0.9, with a balance within the classes, and in a number comparable with the generated fake dataset. We implemented the classifier using Keras, leaning on the Tensorflow back-end, and using categorical accuracy as an evaluation metric. We obtained an accuracy of 70.01%, which is very common to the results of other classifiers in the state-of-the-art [31].

The dataset was divided following the scheme 80-10-10. Due to the unbalancing of the data, every set was balanced to have an equal distribution over the classes on the training set, validation set and test set. Then, we tested the model trained on CelebA and AffectNet DBs with 7, 12 and 17 AUs. The desired accuracy of the results on a fake test set should be close to the accuracy on a real data test set (70.01%). Therefore, we did not aim to reach an accuracy of 100%, but the accuracy closest to the network accuracy on the original test set. The tests are further divided into two other methods: random batch method and selected batch method. These two methods differ only in the methodology of data selection. In the first case, a random batch of images is taken from the selected dataset. In the second case, we manually selected the images with a higher definition and clarity. We added the following constraints to the image sets to ensure consistency in terms of comparison between results on the original and fake test sets:

- The number of images of the fake test sets, both in the random batch method and selected batch method, must be equal to the number of the original test set.
- The emotions must be assigned to the images following the distribution of classes of the original test set.

The accuracy results of the training of the classifier on the resulting images are shown in Table 7. Similarly to the FID test, the training on the AffectNet dataset produced more accurate results than the training on CelebA dataset, and in general, the best results are given by the use of images selected by CelebA with the model trained on AffectNet. The difference between the images of the two datasets is very small because the manually selected images were clear and not very noisy. It is to be noticed that the differences between 7, 12 and 17 AUs are reversed compared to the previous test on FIDs. The emotions were markedly more visible, and therefore the classifier was able to predict the correct classes with higher accuracy.

5.3.1 Interview Test

Finally, we asked human participants to evaluate our model since the *visual examination of samples by human raters is one of the common and most intuitive ways to evaluate GANs* [1, p.30]. We firstly asked participants to assign an emotion to each image, and then, to rate the accuracy, in terms of authenticity, of each image using a scale from 1 (image clearly artefact) to 5 (image indistinguishable from an original). Each participant labelled and rated 30 randomly selected images which were modified by our network. The images have been modified using synthesised vectors with 12 AU, as an intermediate method between 7 AU and 17 AU.

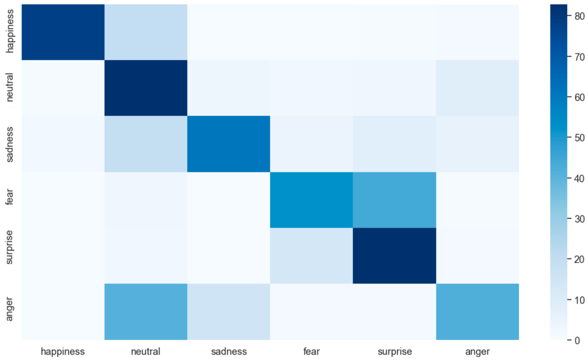


Fig. 7: The confusion matrix of the classification. Darker colours represent higher accuracy.

The test was conducted online without the direct supervision of the experimenter, therefore we decided to use a poll of 30 images to prevent the participants from losing their concentration or interest during the test. We recruited 50 participants, aged between 18 and 60 years old.

The performance of the classification is shown in Figure 7 using a confusion matrix. The Figure shows that the images labelled as Happiness, Neutral, Surprise and Sadness were classified with higher accuracy by the participants, the Fear emotion was classified with lesser accuracy, while images labelled as Anger emotions were misclassified more frequently. In particular, Anger labelled images were often confused as Neutral, and in fewer cases as Sadness. Images labelled as Fear have often been classified as Surprise. We believe that the misclassification of these two emotions is because Fear and Surprise present relevant similarities in the way they are expressed ([6, 9]), and images modified with 12 AU might not be clear enough to be recognised.

Participants also rated the images labelled as showing Happiness (mean value 3.39), Neutral (mean value 3.92) and Sadness (mean value 3.43) emotions as with higher authenticity compared to those labelled with Fear (mean value 2.07), Surprise (mean value 3.06), and Anger (mean value 2.84) emotions. This result indicates that the generation of images showing the emotional state of Fear produces results with a certain level of distortion. We believe that this is due to the difficulty of generating clear images of faces with wide-open mouths or wide eyes.

Table 7: Accuracy of the classifier trained on AffectNet and CelebA datasets using random and manually selected batches. The values in the table represent the accuracy and relative loss of the classifier on the specific dataset according to the two methods.

AffectNet dataset

Images from	7AU	12AU	17AU
AffecteNet random batch	52.52% (3.62)	54.39% (3.22)	57.62% (3.30)
AffectNet selected batch	54.92% (3.54)	61.44% (3.11)	68.92% (2.89)
CelebA random batch	58.97% (3.28)	60.01% (3.05)	61.50% (3.01)
CelebA selected batch	61.02% (3.59)	67.93% (2.89)	69.02% (2.78)

CelebA dataset

AffectNet random batch	48.74% (4.22)	52.23% (4.02)	55.91% (3.82)
AffectNet selected batch	51.92% (3.88)	57.66% (3.81)	63.03% (3.01)
CelebA random batch	55.03% (4.02)	58.27% (3.22)	59.63% (3.34)
CelebA selected batch	60.15% (3.22)	63.03% (3.17)	64.63% (3.03)

5.3.2 Action Unit Test on the Robot

The generated AUs have been used to modify the facial features of a Furhat robot. The robot comes with a set of libraries that allows the manipulation of the robot's characteristics, called gestures. We found a reasonable correspondence between the 17 Action Units and the facial parameters provided by Furhat Robotics[4]. Figure 8 shows examples of facial emotions based on the AUs generated by our approach on the robot.

To investigate people's perception of the generated emotions, we asked 40 participants to classify each emotion as expressed by a robot and one of the people in Figure 4 (i.e., the human emotions are the same as shown in the Figure 4). The sample of participants consisted of 19 males and 21 females, aged between 18 and 69 (avg. 35.27, std. 13.11). Each of them classified the perceived discrete emotions expressed by the robot and human in 50 images, selecting them between the following set [as multiple choice]: Happiness, Surprise, Fear, Anger and Sadness. We also asked participants to select their confidence level for the selection of the showed emotion using a 5-point semantic scale [1 = not at all, and 5 = very much].

The performance of the emotion association as perceived by the participants is shown in Figure 9. The Figure shows that the images labelled as Anger, Sadness and Fear were correctly classified with higher accuracy. Participants, even if they correctly recognised Surprise, were more undecided, and classified Surprise also as Fear. These two emotions are, however, notably difficult to distinguish due to perceptual-attentional limitation, and the context in which they are used [43]. Finally, participants were not able to identify the Happiness emotion of the robot, mistaking it for Surprise.

We believe this has been caused by the impossibility of finding a direct correspondent Furhat's gesture for all AUs (i.e., lip corner puller), and, as a consequence, we chose the closest available gesture. This produced slightly different expressions compared to the ones generated directly for human faces.

We also observed a strong positive correlation between the emotions chosen by the participants and the agent used (i.e., robot or human) with an increase in the Happiness emotion when it was expressed by a human ($p < 0.001 r = 0.1271$), a decrease in the Sadness and Anger emotions when it was expressed by the robot (respectively, $p = 0.014 r = -0.055$, and $p < 0.001 r = -0.140$). We did not find any statistically significant relationship between the type of agent that expressed the emotion and the Surprise and Fear emotions. Indeed, they were both similarly recognised by the participants.

6 Conclusions

This work presents a two-layer architecture capable of modifying the expression of a human face from images by taking into account the need

[4] We used a combination of ARKitParams blendshapes and CharParams facial offsets to animate the synthesised expressions to the robot's face https://docs.furhat.io/facecore

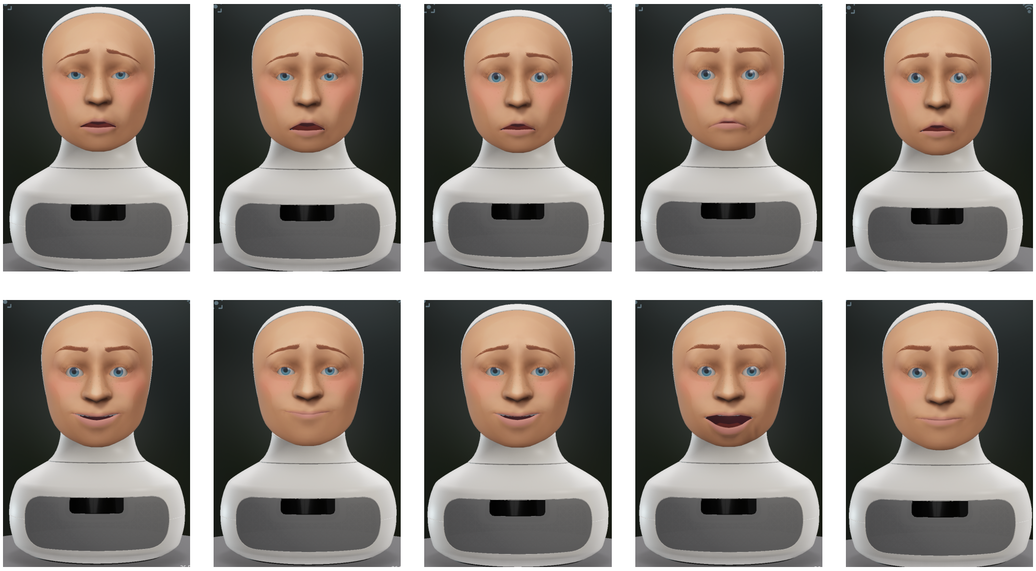


Fig. 8: Examples of the facial expressions for Sadness (top row) and Happiness (lower row) based on the AUs generated by our approach on the robot Furhat.

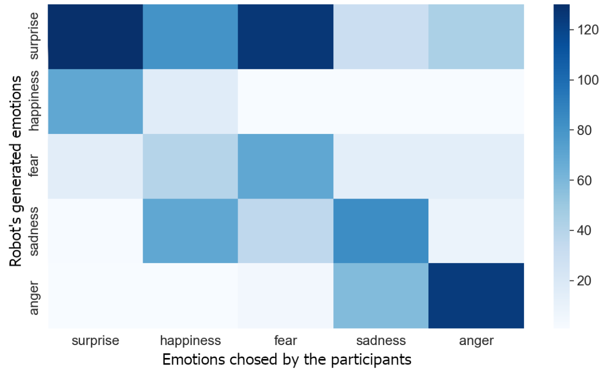


Fig. 9: The confusion matrix of the participants' ranking of the robot's expressions. Darker colours represent higher accuracy.

to provide a relationship between expression and emotion. To obtain a certain variety of expressions for each discrete emotion, we proposed an architecture consisting of two modules, called the AU Generator Module and the GANimation module. In literature, the presence of an architecture like GANimation allows the fine modelling of expressions based on the contribution of studies on the FACS based on Action Units. The original GANimation architecture does not deal with discrete emotions, although it allows transferring the emotion described by a set of Action Units to a specific image providing a considerable variety of expressions. Therefore, an AU Generator module has been developed to generate a set of Action Units that describe a facial expression according to the desired emotion.

The use of the presented module combined with the GANimation architecture allowed us to exploit the characteristics of this architecture with a model of discrete emotions. In particular, the AU Generator module has been developed as a conditional Generative Adversarial Network that can generate different vectors of AU for each discrete emotion and describe different expressions. CGAN architecture allowed to generate data with a high rate of variability and is always consistent with the class on which the generation is conditioned.

We conducted several tests to prove the goodness of the proposed approach, starting with the AU Generator Module trained on AffectNet and FERG DB datasets. These tests showed the goodness of the model to generate realistic data with a satisfactory variety, when compared to the original dataset, especially with the FERG dataset. The model's ability to condition the generated AU vectors to the emotion expressed by the facial expressions they encode was also tested. Finally, the model has been tested in its entirety, including the GANimation module, evaluating the result in terms of final synthetic images. These tests were performed by training the model on CelebA and AffectNet datasets to assess whether the results obtained in the evaluation of the previous module are reflected in the quality of the final images. These tests highlighted the goodness of the model, in terms of photo-realism, evaluated through the use of the FID metrics and conditioning. These tests also showed that the GANimation module benefited from the greater variety and number of elements in the AffectNet dataset compared to the CelebA dataset.

The proposed model achieved satisfactory results, but the quality of the synthetic data and a possible extension to include different emotional models represent a desirable future development. In this work, we used low-resolution images (typically 128x128), due to hardware limitations which could not provide high computational power, but the datasets consisting of high-resolution images generate images with clearer and more evident expressions. Moreover, the variability and naturalness obtained by the combination of the two different GANs could be compared with the generation of facial emotion as an end-to-end process. Finally, the model could also be extended by including the representation of dimensional emotion models to introduce the valence and arousal, or the intensity dimensions [34]. A preliminary test to collect people's perceptions of the generated emotions has been conducted using a Furhat robot. We were not able to directly use the exact corresponding gesture for each AU used, however, participants were able to correctly classify most of the emotions expressed by the robot. These results also outlined the importance of some stimuli, such as the lip corner puller or stretcher, as a very distinctive feature between the expressions, and of the context used to clearly differentiate between known-mistaken emotions (such as surprise and fear.

Declarations

Funding: The research leading to these results received partial funding from the Italian PON I&C 2014-2020 within the BRILLO research project "Bartending Robot for Interactive Long-Lasting Operations", no. F/190066/01-02/X44 and by PON R&I 2014-2020 "Istruzione e ricerca per il recupero – REACT-EU", Azione IV.4 CUP E65F21002920003.

Conflict of Interest: The authors declare that they have no conflict of interest.

Data Availability: No Dataset produced.

References

[1] Pros and cons of gan evaluation measures. *Computer Vision and Image Understanding*, 179:41–65, 2019. ISSN 1077-3142. doi: https://doi.org/10.1016/j.cviu.2018.10.009.

[2] Deepali Aneja, Alex Colburn, Gary Faigin, Linda Shapiro, and Barbara Mones. Modeling stylized character expressions via deep learning. In *Asian Conference on Computer Vision*, pages 136–153. Springer, 2016.

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR, 06–11 Aug 2017.

[4] Casey Bennett and S. Sabanovic. Deriving minimal features for human-like facial expressions in robotic faces. *International Journal of Social Robotics*, 6:367–381, 08 2014. doi: 10.1007/s12369-014-0237-z.

[5] Benedetta Bucci, Alessandra Rossi, and Silvia Rossi. Action unit generation through dimensional emotion recognition from text. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1071–1076, 2022. doi: 10.1109/RO-MAN53752.2022.9900535.

[6] Linda A. Camras. Children's understanding of facial expressions used during conflict encounters. *Child Development*, 51(3):879–885, 1980. ISSN 00093920, 14678624.

[7] Pierluigi Carcagnì, Marco Del Coco, Marco Leo, and Cosimo Distante. Facial expression recognition and histograms of oriented gradients: a comprehensive study.

SpringerPlus, 4:645, 11 2015. doi: 10.1186/s40064-015-1427-3.

[8] Filippo Cavallo, Francesco Semeraro, Laura Fiorini, Gergely Magyar, Peter Sinčák, and Paolo Dario. Emotion modelling for social robotics applications: A review. *Journal of Bionic Engineering*, 15(2):185–203, 2018.

[9] Justin Chamberland, Annie Roy-Charland, Melanie Perron, and Joël Dickinson. Distinction between fear and surprise: An interpretation-independent test of the perceptual-attentional limitation hypothesis. *Social Neuroscience*, 12, 10 2016. doi: 10.1080/17470919.2016.1251964.

[10] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018. doi: 10.1109/CVPR.2018.00916.

[11] Nikhil Churamani, Pablo Barros, Erik Strahl, and Stefan Wermter. Learning empathy-driven emotion expressions using affective modulations. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2018. doi: 10.1109/IJCNN.2018.8489158.

[12] Jia Deng, Gaoyang Pang, Zhiyu Zhang, Zhibo Pang, Huayong Yang, and Geng Yang. cgan based facial expression recognition for human-robot interaction. *IEEE Access*, 7:9848–9859, 2019. doi: 10.1109/ACCESS.2019.2891668.

[13] Paul Ekman. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200, 1992. doi: 10.1080/02699939208411068.

[14] Paul Ekman and Wallace V Friesen. Measuring facial movement. *Environmental psychology and nonverbal behavior*, 1(1):56–75, 1976.

[15] Paul Ekman, Wallace V Freisen, and Sonia Ancoli. Facial signs of emotional experience. *Journal of personality and social psychology*, 39(6):1125, 1980.

[16] Diego R. Faria, Mario Vieira, Fernanda C.C. Faria, and Cristiano Premebida. Affective facial expressions recognition for human-robot interaction. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 805–810, 2017. doi: 10.1109/ROMAN.2017.8172395.

[17] Chiara Filippini, David Perpetuini, Daniela Cardone, Antonio Maria Chiarelli, and Arcangelo Merla. Thermal infrared imaging-based affective computing and its application to facilitate human robot interaction: A review. *Applied Sciences*, 10(8), 2020. ISSN 2076-3417. doi: 10.3390/app10082924.

[18] Z. Geng, C. Cao, and S. Tulyakov. 3d guided fine-grained face manipulation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9813–9822, Los Alamitos, CA, USA, jun 2019. IEEE Computer Society. doi: 10.1109/CVPR.2019.01005.

[19] Rachel Gockley, Reid Simmons, and Jodi Forlizzi. Modeling affect in socially interactive robots. In *ROMAN 2006 - The 15th IEEE International Symposium on Robot and Human Interactive Communication*, pages 558–563, 2006. doi: 10.1109/ROMAN.2006.314448.

[20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

[22] Ahmad Hesam, Sofia Vallecorsa, Gulrukh Khattak, and Federico Carminati. Evaluating power architecture for distributed training of generative adversarial networks. In Michèle Weiland, Guido Juckeland, Sadaf Alam, and Heike Jagode, editors, *High Performance Computing*, pages 432–440, Cham, 2019. Springer International Publishing. ISBN 978-3-030-34356-9.

[23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[24] Ruud Hortensius, Felix Hekele, and Emily S. Cross. The perception of emotion in artificial agents. *IEEE Transactions on Cognitive and Developmental Systems*, 10(4):852–864, 2018. doi: 10.1109/TCDS.2018.2826921.

[25] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[26] Dae-Kwan Ko, Dong-Han Lee, and Soo-Chul Lim. Continuous image generation from low-update-rate images and physical sensors through a conditional gan for robot teleoperation. *IEEE Transactions on Industrial Informatics*, 17(3):1978–1986, 2021. doi: 10.1109/TII.2020.2991764.

[27] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[28] Ohman A. Lundqvist D., Flykt A. The karolinska directed emotional faces - kdef. *Karolinska Institutet*, CD ROM from Department of Clinical Neuroscience, Psychology section, 1998.

[29] X. Mao, Q. Li, H. Xie, R. K. Lau, Z. Wang, and S. Smolley. Least squares generative adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2813–2821, Los Alamitos, CA, USA, oct 2017. IEEE Computer Society. doi: 10.1109/ICCV.2017.304.

[30] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014.

[31] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, Jan 2019. ISSN 2371-9850. doi: 10.1109/taffc.2017.2740923.

[32] Duc Thanh Nguyen, Wanqing Li, and Philip O. Ogunbona. Human detection from images and videos: A survey. *Pattern Recognition*, 51:148–175, 2016. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2015.08.027.

[33] Aäron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. Conditional image generation with pixelcnn decoders. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4797–4805, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

[34] Jonathan Posner, James A. Russell, and Bradley S. Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, 17(3):715–734, 2005. doi: 10.1017/S0954579405050340.

[35] Albert Pumarola, Antonio Agudo, Aleix M. Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 835–851, Cham, 2018. Springer International Publishing.

[36] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[37] Suman Ravuri, Shakir Mohamed, Mihaela Rosca, and Oriol Vinyals. Learning implicit generative models with the method of learned moments. In *International Conference on Machine Learning*, pages 4314–4323. PMLR, 2018.

[38] Niyati Rawal and Ruth Maria Stock-Homburg. Facial emotion expressions in human–robot interaction: A survey. *International Journal of Social Robotics*. doi: 10.1007/s12369-022-00867-0.

[39] Alessandra Rossi, Kerstin Dautenhahn, Kheng Lee Koay, and Michael L. Walters. How social robots influence people's trust in critical situations. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1020–1025, 2020. doi: 10.1109/RO-MAN47096.2020.9223471.

[40] S. Rossi and M. Ruocco. Better alone than in bad company: Effects of incoherent non-verbal emotional cues for a humanoid robot. *Interaction Studies*, 20(3):487–508, 2019. doi: 10.1075/is.18066.ros.

[41] S. Rossi, M. Larafa, and M. Ruocco. Emotional and behavioural distraction by a social robot for children anxiety reduction during vaccination. *International Journal of Social Robotics*, 12(3):765–777, 2020.

[42] Silvia Rossi, Alessandra Rossi, and Kerstin Dautenhahn. The secret life of robots: Perspectives and challenges for robot's behaviours during non-interactive tasks. *International Journal of Social Robotics*, 2020. doi: 10.1007/s12369-020-00650-z.

[43] Annie Roy-Charland, Melanie Perron, Olivia Beaudry, and Kaylee Eady. Confusion of fear and surprise: A test of the perceptual-attentional limitation hypothesis with eye movement monitoring. *Cognition emotion*, 28, 01 2014. doi: 10.1080/02699931.2013.878687.

[44] James A. Russell, Jo-Anne Bachorowski, and José-Miguel Fernández-Dols. Facial and vocal expressions of emotion. *Annual Review of Psychology*, 54(1):329–349, 2003. doi: 10.1146/annurev.psych.54.101601.145102.

[45] Mingyang Shao, Silas Franco Dos Reis Alves, Omar Ismail, Xinyi Zhang, Goldie Nejat, and Beno Benhabib. You are doing great! only one rep left: An affect-aware social robot for exercising. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 3811–3817, 2019. doi: 10.1109/SMC.2019.8914198.

[46] Lingxiao Song, Zhihe Lu, Ran He, Zhenan Sun, and Tieniu Tan. Geometry guided adversarial facial expression synthesis. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM '18, page 627–635, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356657. doi: 10.1145/3240508.3240612.

[47] Matteo Spezialetti, Giuseppe Placidi, and Silvia Rossi. Emotion recognition for human-robot interaction: Recent advances and future perspectives. *Frontiers in Robotics and AI*, 7:145, 2020. ISSN 2296-9144. doi: 10.3389/frobt.2020.532279.

[48] Hao Tang, Wei Wang, Songsong Wu, Xinya Chen, Dan Xu, Nicu Sebe, and Yan Yan. Expression conditional gan for facial expression-to-expression translation. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 4449–4453, 2019. doi: 10.1109/ICIP.2019.8803654.

[49] Md. Zia Uddin, Mohammad Mehedi Hassan, Ahmad Almogren, Atif Alamri, Majed Alrubaian, and Giancarlo Fortino. Facial expression recognition utilizing local direction-based robust features and deep belief network. *IEEE Access*, 5:4525–4536, 2017. doi: 10.1109/ACCESS.2017.2676238.

[50] Thomas Unterthiner, Bernhard Nessler, Calvin Seward, Günter Klambauer, Martin Heusel, Hubert Ramsauer, and Sepp Hochreiter. Coulomb GANs: Provably Optimal Nash Equilibria via Potential Fields. *arXiv e-prints*, art. arXiv:1708.08819, August 2017.

[51] Paul Viola, Michael J Jones, and Daniel Snow. Detecting pedestrians using patterns of motion and appearance. *International Journal of Computer Vision*, 63(2):153–161, 2005.

[52] Michael L. Walters, Dag S. Syrdal, Kerstin Dautenhahn, René te Boekhorst, and Kheng Lee Koay. Avoiding the uncanny valley: robot appearance, personality and consistency of behavior in an attention-seeking home scenario for a robot companion. *Autonomous Robots*, 24(2):159–178, Feb 2008. ISSN 1573-7527. doi: 10.1007/s10514-007-9058-3.

[53] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, September 2018.

[54] Siyue Xie and Haifeng Hu. Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks. *IEEE Transactions on Multimedia*, 21(1):211–220, 2019. doi: 10.1109/TMM.2018.2844085.

[55] Chuang Yu and Adriana Tapus. Interactive robot learning for multimodal emotion recognition. In Miguel A. Salichs, Shuzhi Sam Ge, Emilia Ivanova Barakova, John-John Cabibihan, Alan R. Wagner, Álvaro Castro-González, and Hongsheng He, editors, *Social Robotics*, pages 633–642, Cham, 2019. Springer International Publishing.