Neurocomputing 528 (2023) 200–216

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Triple-BigGAN: Semi-supervised generative adversarial networks for image synthesis and classification on sexual facial expression recognition

Abhishek Gangwar [a,c,*], Víctor González-Castro [a,b], Enrique Alegre [a,b], Eduardo Fidalgo [a,b]

[a] Department of Electrical, Systems and Automatic Engineering, Universidad de León, Campus de Vegazana s/n, 24071 León, Spain
[b] Researcher at INCIBE (Spanish National Cybersecurity Institute), León, Spain
[c] Centre for Development of Advanced Computing (CDAC), Mumbai 400049, India

ARTICLE INFO

Article history:
Received 27 April 2022
Revised 13 December 2022
Accepted 8 January 2023
Available online 12 January 2023

Keywords:
Facial expressions
Pornography
Not safe for work (NSFW)
Obscene image retrieval
Deep learning
Emotion detection

ABSTRACT

Automatic recognition of facial images showing erotic expressions can help to understand our social interaction and to detect non-appropriate images even when there is no nakedness present in them. This paper contemplates, for the first time, to exploit facial cues applied to automatic Sexual Facial Expression Recognition (SFER). With this goal, we introduce a new dataset named Sexual Expression and Activity Faces (SEA-Faces-30k) for SFER, which contains 30k manually labeled images under three categories: erotic, suggestive-erotic, and non-erotic. Deep Convolutional Neural Networks require large-scale annotated image datasets with diversity and variations to be properly trained. Unfortunately, gathering such a massive amount of data is not feasible in this area. Therefore, we present a new semi-supervised GAN framework named Triple-BigGAN, which learns a generative model and a classifier simultaneously. It learns both tasks in an end-to-end fashion while using unlabeled or partially labeled data. The Triple-BigGAN framework shows promising classification performance for the SFER task (i.e., 93.59%) and other five benchmark datasets, i.e., FER-2013, CIFAR-10, Expression in-the-Wild (ExpW), Modified National Institute of Standards and Technology database (MNIST), and Street View House Numbers (SVHN). Next, we evaluated the quality of samples generated by Triple-BigGAN with a resolution of 256×256 pixels using Inception Score (IS) and Frechet Inception Distance (FID). Our approach obtained the best FID (i.e., 19.94%) and IS (i.e., 97.98%) scores on the SEA-Faces-30k dataset. Further, we empirically demonstrated that synthetic erotic face images generated by Triple-BigGAN could also help in improving the classification performance of deep supervised networks.

1. Introduction

Automatic facial expression is an active field with multiple applications, from video surveillance to emotion-based photo capturing and tagging [78,6,1,89,28]. In the research related to facial expression recognition, the existing annotated datasets and state-of-the-art methods mostly cover the six discrete emotions proposed by Ekman [26]: anger, disgust, fear, happiness, sadness, and surprise. Ekman and Friesen stated that these basic facial expressions are common to all humans, irrespective of their birth country or culture. Richard and Bernice [44] proposed 15 emotions by extending the list with emotions such as anger, anxiety, compassion, aesthetic experience, depression, guilt, envy, fright, gratitude, hope, jealousy, love, happiness, pride, relief, sadness, and shame. In [24], Ekman suggested expanding the list of basic emotions to 17 by adding excitement, amusement, contempt, embarrassment, pride in achievement, relief, satisfaction, contentment, sensory pleasure, guilt, and shame. These normal facial expressions that last between 0.5 to 4 s are also called macro expressions. Apart from macro-expressions, researchers have also explored micro-expressions, i.e. that last less than half a second [47].

This paper focuses on detecting sexual expressions as macro expressions on human faces using automatic Sexual Facial Expression Recognition (SFER), which would make it possible to verify if an image has erotic nature. Nowadays, in digital forensics, Law Enforcement Agencies (LEAs) use Multimedia-Forensic Analysis Tools and Facial-Forensic Analysis Tools to deal with the growing volume of data seized from cybercrimes [72,73]. Crimes, where Child Sexual Exploitation Material (CSEM) is involved, are

* Corresponding author at: Department of Electrical, Systems and Automatic Engineering, Universidad de León, Campus de Vegazana s/n, 24071 León, Spain.
E-mail addresses: abhishekg@cdac.in (A. Gangwar), victor.gonzalez@unileon.es (V. González-Castro), ealeg@unileon.es (E. Alegre), efidf@unileon.es (E. Fidalgo).

https://doi.org/10.1016/j.neucom.2023.01.027
0925-2312/

especially sensitive, due to the nature of the handled data. One of the strategies that are used to detect CSEM automatically [30] is to detect images that contain pornography and to estimate the age of people in such images. The goal is to find out obscene images where minors are present, which would be categorized as possible CSEM. Nevertheless, some images contain only faces (e.g. in close-up photos), making the retrieval of such material a challenging task for pornography and, subsequently, CSEM detectors. When such images are present, the SFER may improve the performance of pornography detectors and, therefore, enrich the Facial-Forensic Analysis.

Deep neural networks have yielded dramatic performance gains in recent years on Computer Vision tasks [32,80,36]. However, these successes are heavily dependent on large training sets of manually annotated data [5,31]. As far as we know, there are not any publicly available large collections of labeled data suitable for training a deep learning model for sexual facial expression detection. Hence, first, we created a manually labeled dataset containing erotic, suggestive-erotic, and non-erotic facial images. We observed that the labelling procedure for the task of sexual facial expression is much more difficult and time-consuming than normal image labelling. Motivated by the aforementioned issues, in this paper we propose a Semi-Supervised Learning (SSL) framework based on Generative Adversarial Networks (GANs) [32], which can learn image representation from data from which only a small part is labeled.

In regard to learning better representations, researchers have been exploiting different methods to utilize unlabeled or partially labeled data for many years [82]. The reason is that the network can learn embedded informative patterns hidden in the data, and then this learning can be transferred to the classifiers, which are trained on the available limited labeled data. That way, such classifiers can then generalize better.

Recently, GANs have achieved impressive success for various types of computer vision problems such as image synthesis [51], style transfer [84], image super-resolution [71], and classification [34]. In general, conventional GANs are specific neural networks in which the training is performed under an unsupervised setting. Their main goal is to generate synthetic samples with a data distribution similar to the input data distribution. During the training of GANs, an adversarial objective is set between a discriminator network and a generator network. The discriminator performs the task of detecting whether the input sample is drawn from the true data or the fake sample synthesized by the generator. The objective of the generator is set to synthesize images that look as if drawn from actual data to the discriminator. The adversarial learning and a competitive game between the discriminator and generator help in protecting the discriminator from over-fitting on the input data, especially when the training data size is small. Finally, the synthetic images generated by the generator can be utilized for various purposes, including data augmentation for improved training of classifiers [19].

One interesting extension of GAN is Conditional GAN (CGAN) [56] where a condition variable can control the generated image. In an alternative approach proposed in [65], the authors build auxiliary classifier GANs (AC-GANs), where the side information is reconstructed by the discriminator instead. Irrespective of the specific approach, this line of research focuses on the supervised setting, where it is assumed that all the images have attribute tags.

Further, the GAN models have been used with semi-supervised learning [82,64,15,86,34]. Also, [87,82] used GANs to perform semi-supervised classification by using a generator-discriminator pair to learn an unconditional model of the data and fine-tune the discriminator using the small amount of labeled data for prediction. Given that labeled data is expensive, it is interesting to explore semi-supervised settings where only a small fraction of the images have class labels. In contrast, a majority of the images are unlabeled.

The key contributions of this paper are summarized as follows:

- We present a novel end-to-end semi-supervised GAN framework named Triple-BigGAN. It is capable of (i) learning a discriminative classifier, as well as (ii) generating high-quality synthesized images from partially labeled data.
- We introduce the task of Sexual Facial Expression Recognition (SFER). It consists on detecting automatically whether a face is showing an expression related to sexuality, either explicit or suggesting. To the best of our knowledge, this is the first work in which this task is tackled.
- We introduce a new image dataset, named SEA-Faces-30k, with 30k manually-labeled facial images. This is the first dataset of images of Sexual Expression and Activity Faces. It can be accessed through our website (https://gvis.unileon.es/dataset/sea-faces/) upon request for research purposes only.
- We empirically demonstrate that Triple-BigGAN provides state-of-the-art classification accuracy on the FER-2013 [1], ExpW [91], MNIST [45], CIFAR-10 [43], SVHN [62] and SEA-Faces-30k datasets. We also show that Triple-BigGAN provides high-quality and high-resolution synthetic samples. Furthermore, we show that adding images generated by Triple-BigGAN to a dataset improves the accuracy of supervised learning-based methods for the task of SFER.

The rest of the paper is organized as follows: First, a revision of works related to ours is addressed in Section 2. In Section 3, we introduce the SEA-Faces-30k dataset. Then, our proposed approach is described in Section 4. The description of the experiments, their results, and a discussion are covered in Section 5 and, finally, Section 6 includes the main findings of our work.

2. Related work

2.1. Related work in GANs

The Triple-BigGAN model proposed in this work has been designed as a GAN framework for joint-distribution matching. There are several extensions of GANs, like Conditional GAN (CGAN) [56], in which a condition variable controls the generation of the images. Numerous CGANs have been introduced in the literature to condition the image generation on class labels [56], images [39], and object/image attributes [68].

Researchers have explored different ways to convert standard GAN into CGAN [56,82,68,23,65,58]. The basic type of CGANs requires supervised information related to the condition variable (s). Springerberg [82] replaced the binary discriminator in standard GAN with a multi-class classifier and presented categorical generative adversarial networks (CatGAN). He trained the generator and the discriminator using information theoretical learning on unlabeled data [82]. Dumoulin et al. [23] and de Vries et al. [16] presented a modified class conditioning in the input to the generator by means of class conditional gains and biases in Batch Normalization layers [38]. In the work carried out by Odena et al. [65], the noise vector input in the standard generator is substituted by a noise vector concatenated with a 1-hot class vector. The objective is to boost conditional samples to maximize the respective class probability predicted with the help of an auxiliary classifier. In [58], the authors modified the discriminator and utilized cosine distance between its features and a set of learned class embeddings

[1] https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge

to provide extra supervision to discriminate between the real and the generated samples. It resulted in the generation of samples in which the features are closer to a learned class prototype.

Some authors have also explored completely unsupervised methodologies to generate samples of a specific type as an alternative to the control variable-based conditional image generation. The authors in [12] modified the input in the standard generator by introducing a latent code vector jointly with the noise vector. The latent codes are then learned by variational mutual information maximization between the latent code and the generator sample in an unsupervised manner. The Adversarially Learned Inference (ALI) [22] method, extended the standard generator, i.e., an encoder, with an additional decoder network. The decoder takes a data sample as input and outputs a synthetic latent vector. The objective of the discriminator is also modified, and it now takes joint pairs – i.e., the latent vector and the data sample – and makes the classification if the pair belongs to an encoder or decoder. The training of the encoder and the decoder modules is performed together to learn the discriminator. In another work, in BiGAN, or Bidirectional GAN [21], the authors introduced an encoder module along with a discriminator and generator in the GAN. The encoder module learns a mapping from data to latent representations. Then, in addition to classifying a real sample rather than a generated sample, the discriminator also discriminates between the encoder's learned representation and the latent space.

Triple-GAN [46] also employs the idea of the conditional generator, but uses adversarial cost to match the two model-defined factorizations of the joint distribution with the one defined by paired data. In addition, Triple-GAN introduced an additional player, i.e., a classifier, in the standard GAN, containing a discriminator, and a generator, to do semi-supervised learning with compatible utilities. In another more recent work, Haque proposed the model External Classifier GAN (EC-GAN) [34], comprising a generator, a discriminator, and a classifier. The EC-GAN trains the classifier in an end-to-end manner along with the discriminator and the generator, however, the major goal of EC-GAN is to utilize synthetic samples generated by the generator to augment the training data for the classifier. EC-GAN did not utilize the pseudo-labels generated by the classifier to improve the training of the generator or the discriminator.

Furthermore, various methods and model architectures have been proposed to enhance and stabilize the training of GANs while generating both high-resolution (i.e., large) and high-quality images. In [40], Karras et al. presented a new training methodology and improvements in the discriminator and generator to generate high-resolution (e.g. 1024×1024) realistic samples. They adopted a progressive training strategy in which generator and discriminator networks with lesser layers are trained on low-resolution images, such as 4×4 pixels in the beginning. Then, incrementally, they kept adding blocks of layers which allowed growing the size of the output in the generator and the size of the input to the discriminator. The step-by-step incriminating of the networks continued until the desired image resolution is obtained.

Another approach, the Style Generative Adversarial Network or StyleGAN [41], extended the progressive GAN and explored multiple improvements to the generator part while keeping the same discriminator and loss functions. They removed the traditional latent vector input layer in the generator and replaced it with a non-linear mapping network (i.e., an 8-layer Multilayer Perceptron (MLP)) which maps a latent code to an intermediate latent space. The intermediate latent space is then used to guide the style at each point in the generator through a new layer called "Adaptive Instance Normalization" (AdaIN). The authors also included another reference of randomness in the form of noise affixed to the whole feature maps after each convolution layer to introduce stochastic variation in the synthetic images. Similar to StyleGAN, Brock et al. presented the BigGAN model [8]. It is another ground-breaking GAN model designing and training strategy to generate high resolution 512×512 realistic images with high quality. More details about BigGAN are provided in Section 4.1.3.

2.2. Related work in facial expression recognition

Existing approaches for Facial Expression Recognition (FER) can be divided into two categories. On the one hand, methods that extract features from a facial image, and use the encoded spatial information for expression classification among seven classes: the six basic emotions proposed by [26] (i.e., anger, disgust, fear, happiness, sadness, and surprise), and a neutral one [63,93,81,78,28]. The other types of approaches for FER [92] involve the use of the Facial Action Coding System (FACS), which describes facial muscle movement using 44 different Action Units (AU) [25]. Each AU corresponds to a specific facial substructure, and the six basic emotions can be categorized by combining multiple AUs. The Emotional Facial Action Coding System is a subset of FACS, which considers only the relevant AUs responsible for such expressions.

Most traditional FER methods are based on hand-crafted image descriptors such as Local Binary Patterns (LBP) [78], Scale Invariant Feature Transform (SIFT) [52] or Histogram of Oriented Gradients (HOG) [66] followed by a classifier such as Support Vector Machine (SVM) [14], Decision Trees (DT) [75] or Artificial Neural Networks (ANN) [18]. State-of-the-art approaches for FER are based on Deep Learning, especially on Convolutional Neural Networks (CNN) [1,89,28,81].

The hand-crafted features give good accuracy in a constrained environment, such that the subject pose expression under fixed head pose and lighting conditions are also stable. However, a significant accuracy drop happens when there is no control over the illumination and head pose angle.

Recently, deep neural networks have been employed to increase the robustness of FER in real-world scenarios. However, the learned deep representations used for FER are often influenced by large variations in individual facial attributes such as ethnicity, gender, or the age of subjects involved in training. The major limitation of this methodology is that it reduces the generalization of the model on unknown identities. Despite noticeable research in the field, modeling inter-subject differences in FER is still persisting as an open challenge.

Following this, various techniques [48,10] have been proposed in the literature to increase the discriminative power of extracted features for FER by increasing the inter-class differences and reducing intra-class variations. More recently, Identity-Aware CNN (IACNN) [55] was presented, and to reduce individual identity-specific information, the authors exploited expression-sensitive contrastive loss and an identity-sensitive contrastive loss. However, it is also reported that the influence of contrastive loss is compromised by large data expansion, and that happens because in contrastive learning the training data is provided in the form of image pairs [10].

Cai et al. [10] presented an Identity-free conditional Generative Adversarial Network (IF-GAN) to minimize the impact of identity-related information by generating a synthetic sample having a facial expression similar to the input sample. This resultant synthetic sample is then utilized for FER to minimize the impact of subject-level variations in the data. However, such a scheme has a challenge, which is that since FER is based on synthetic data, its performance is influenced not only by the quality of the generated data but also by the performance of the expression transfer between the input sample to the synthetic sample. In [88], authors

proposed De-expression Residue Learning (DeRL) to learn subject-independent facial expression representations.

More recently, StarGAN [13] was presented to edit facial expressions and attributes. It is a multi-domain approach that learns the generation of facial expressions and the transfer of facial attributes simultaneously. The system has been designed to control the target facial expression according to the facial expression fed along with the input face to edit.

As an extension to prior work, [20] introduced ExprGAN, a facial expression editing GAN which can learn the potency of the facial emotion by exploiting special encoding of the expression label. They do not require intensity level values; however, various desired expression styles can be generated. The intensity of synthesized emotion can also be controlled from low to high through an expression controller module. However, the approach is not capable enough to generate facial emotions such as compound expressions. Pumarola et al. [69] presented a system based on the coupling GAN and Action Units (AUs) to synthesize facial emotions drawn to form a more extensive dataset continuously. Nevertheless, the approach requires a large amount of labeled data especially, AUs.

Success in various computer vision classification problems relies heavily on the availability of annotated datasets. Literature related to FER presents a significant number of publicly available datasets, summarised in Table 1. However, to the best of our knowledge, none of the existing datasets contain facial images with sexual expressions.

In this work, we propose the Sexual Expression and Activity Faces (SEA-Faces-30k) dataset, the first publicly available dataset related to erotic facial images.

2.3. Related work in sexual facial expression recognition

Even though the automatic analysis of sexual expressions in faces has significant importance, this subject has been explored very little so far (probably due to the sensitive nature of this domain). The study carried out by Rosemary Basson [3] is still considered the most exhaustive observational study of facial expressions of sexual excitement. This analysis is based on data collected from 382 women and 312 men and through 10000 cycles of sexual arousal and orgasm. They analyzed some common behavior during sexual activity contraction of the musculature surrounding the mouth, the opening of the mouth, clenched jaws, or flared nostrils, among others. Fernández-Dols et al. [29] observed the facial expressions in 100 video clips containing an episode of sexual excitement that concluded in an orgasm by volunteers. They coded the facial regions using FACS, and reported that there were nine combinations of muscular movements produced by at least 5% of the video senders. These combinations were consistent with facial expressions of sexual excitement described in [3].

The aforementioned researches studied the correlation between sexual activities and facial expressions. However, in our work, our goal is to use the facial region as a global feature and to analyse if erotic facial images – i.e., those related to sexual activities – can be discriminated from non-erotic facial images from people without any sexual activity.

3. Sexual Expression and Activity Faces Dataset (SEA-Faces-30k)

The SEA-Faces-30k dataset has three categories based on the intensity of eroticism in the facial region: erotic, suggestive-erotic, and non-erotic. Figs. 1–3 depicts some examples of the erotic, suggestive-erotic, and non-erotic categories, respectively.

Table 1
A summary of publicly available datasets for FER.

Dataset	Num. images/ videos	Num. Expressions
CK+ [53]	593 images	6 basic + neutral + contempt
FER-2013	35,887 images	6 basic + neutral
MMI [67]	740 images, 2900 videos	6 basic + neutral
Multi-PIE[33]	755,370 images	6 basic
EmotioNet [5]	1 M images	23 basic or compound
AffectNet [60]	450 K images	6 basic + neutral
ExpW [91]	91,793 images	6 basic + neutral

https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge.

3.1. Dataset statistics

SEA-Faces-30k contains 30,817 faces from the images collected from the Internet and the faces are annotated manually with sexual expressions. Face information changes depending on age, sex, face shape, or skin colour, significantly impacting facial expression analysis. To control these changes and to have enough diversity in data, images in SEA-Faces-30k have variations in terms of age, ethnicity, gender, expression, scene complexity, sexual activity, illumination, head orientation, image resolution, and artifacts on the face such as glasses, hats, beards, or jewelry. Apart from these challenges, sometimes it has been observed that seized images related to pornography or CSEM have low resolution and, hence, in SEA-Faces-30k, we also kept some low-resolution images in the dataset. Among the gender distribution, around 80% of faces in the dataset belong to females and 20% are from males. We found it very hard to obtain a large number of male faces with prominent sexual expressions. The main features of SEA-Faces-30k are summarized in Table 2.

3.2. Dataset creation

SEA-Faces-30k has been created through a multi-stage process: (i) data crawling from the Internet, (ii) removal of near duplicated images, (iii) face detection and alignment, and (iv) manual filtering and categorization. Each step is described in detail in the next paragraphs.

Stage 1: Data Crawling

We performed two types of data crawling to collect images in our dataset. First, we scrapped around 50k pornographic images from two popular pornographic websites. We found that the downloaded images have not only varied content, e.g. from a single-clothed model posing to nude group sexual activity, but also some images were non-pornographic. To ensure the pornographic context in the erotic category of facial images, we utilized the open-source Yahoo open not suitable for work (NSFW) pornography detector[2], which assigns a score to each image depending on the pornographic content.

Then, using the pornographic score obtained from NSFW, we selected the high-score images for the erotic category, the medium-score images for the suggestive-erotic category, and the remaining ones for the non-erotic category images for further steps. Specifically, the crawled images were divided as: *erotic* (score $\in [0.85, 1.0]$), *suggestive-erotic* (score $\in (0.35, 0.85)$) and *non-erotic* (score $\in [0, 0.35]$). The face cropping and possible classification errors were corrected manually in further stages.

We noticed that in the crawled data some persons were featured in multiple images and also the majority of the images were captured in a controlled environment. Therefore, in order to have more variability and general web (i.e., wild) images in the data,

[2] https://github.com/yahoo/open_nsfw

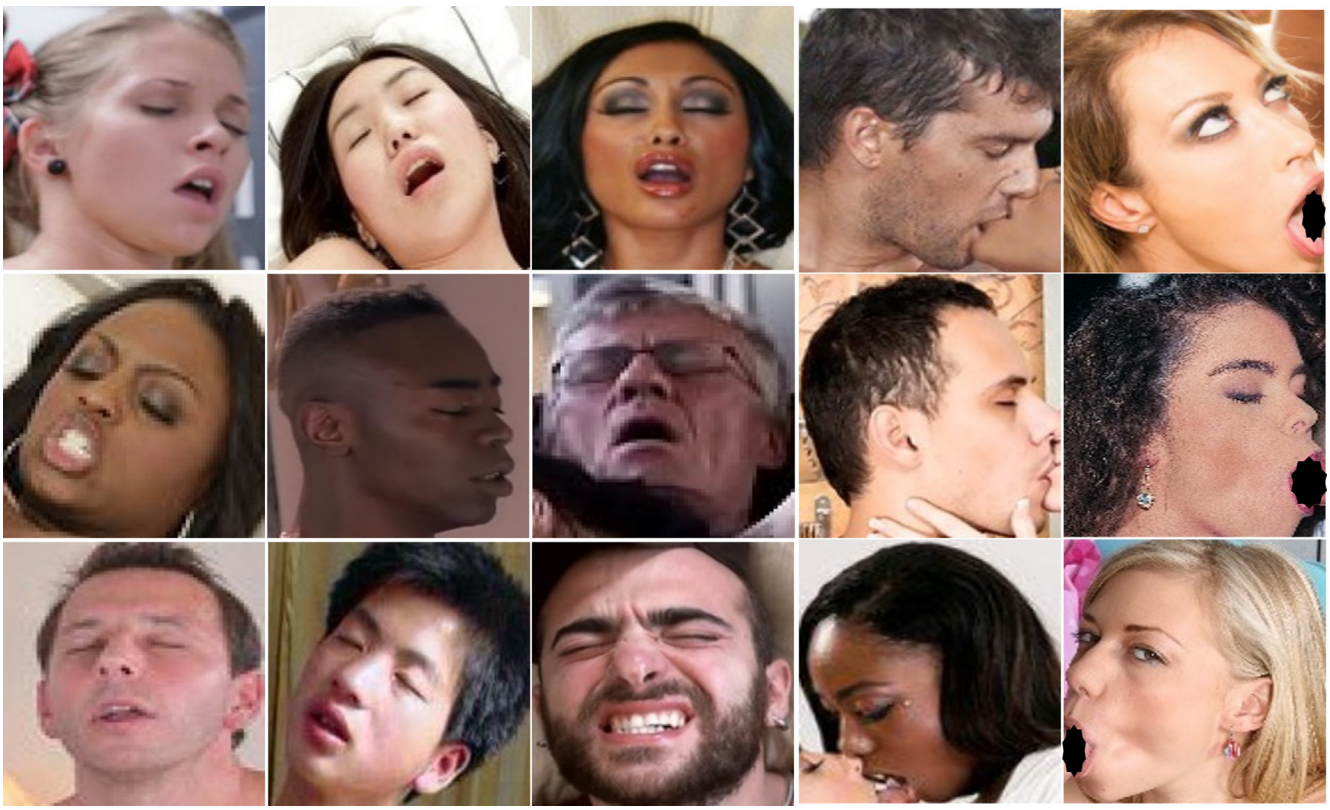

Fig. 1. Examples of images from the class "erotic" of SEA-Faces-30k.

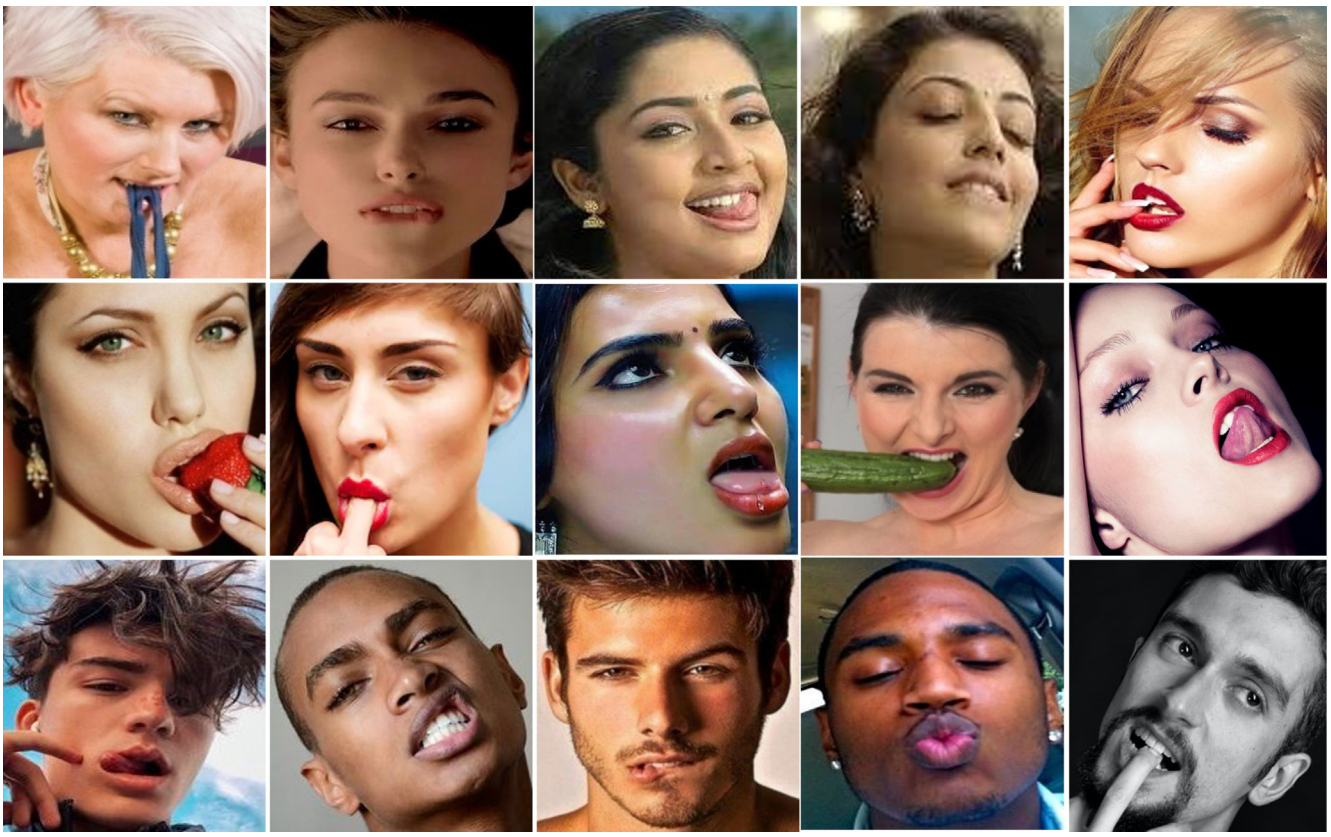

Fig. 2. Examples of images from the class "suggestive-erotic" in SEA-Faces-30k.

Fig. 3. Examples of images from the class "non-erotic" in SEA-Faces-30k.

Table 2
Categories and number of images for each category in the SEA-Faces-30k dataset.

Category	Num. images
Erotic	10399
Suggestive-Erotic	10160
Non-Erotic	10258
Total Images	30817

additionally, we used the Google image search engine to retrieve 10k additional images in each of the three categories. For the *erotic* class, we used the following search keywords: "fellatio", "blow job", "cunnilingus", "anilingus", "cum facial", "orgasm", "pussy-licking", "kissing", "fucking", "cum shot" and "masturbation". For the *suggestive-erotic* class, we used the words: "sexy model", "sexy face", "genital posing", "sexual posing", "sexy lady", "erotic face", "porn stars", and "horny". Finally, to enhance the *non-erotic* data with challenging samples, we added images crawled with the following queries: "mouth open", "human pose", "men with pose", "girls with pose", "women with pose", "happy face", "crying face", "men playing sports", "women playing sports", "girls playing sports", "face covering", "surprise", "men in pain", "boys in pain", "women in pain".

Through the strategies mentioned above, we captured the context of the pornographic nature in the facial expressions in erotic categories and also obtained diverse face variations in the data. Finally, we collected 80,172 coarsely labeled images under three categories in this stage for further processing.

Stage 2: Removal of Near Duplicate Samples

We observed that in the crawled data, there were many images with different versions of the same image, e.g. the same image with different resolutions or with different names. There were also images versions with minor variations. To remove such near duplicated images, we used a perceptual hashing [7] method, i.e. pHash[3] to delete all the images with a dis-similarity score, i.e. Hamming distance, lower than four. The hamming distance of the pHash codes with 64 bits length ranges between 0 and 64, with 0 meaning completely similar and towards 64 representing more dissimilarity. We empirically found that a threshold of four is good enough to detect duplicate or near-duplicate images in our data.

Stage 3: Face Detection and Alignment

Next, in the step of *Face detection and alignment*, we used the RetinaFace detector[4], for detecting the faces on the images, as well as for getting the following landmark points: left and right eye centers, left and right mouth corners, and nose tip. These landmark points are used for face alignment. We only selected the face images with a resolution higher than 50×50 pixels and the selected faces were re-scaled to 256×256 pixels.

Stage 4: Manual Filtering and Categorization

So far, we have obtained three subsets containing facial images coarsely labeled as "erotic", "suggestive-erotic", and "non-erotic", respectively. Then, we carried out the manual verification and categorization of the faces in each subset (i.e., the facial images belonging to each category were selected through visual inspection). During this step, we also removed all the images in which either a real human face was not present due to an error of the face detector or cartoon/statue faces. Moreover, we removed faces with low quality, occlusions, large poses, and overlapping.

To decide the correct category during manual labeling, we first fixed the understanding of each category based on the following criteria:

- The *erotic* category should either include faces with expressions

[3] https://www.phash.org
[4] https://github.com/deepinsight/insightface/tree/master/RetinaFace

made during orgasms or faces carrying out oral sexual activities. Specifically, the faces in this category should be the facial region images that can cause arousal in humans.
- The *suggestive-erotic* category faces are those which, by seeing them, people should feel the sexual intent in these faces. Nevertheless, the erotic intensity in these images is not as high as mentioned in the erotic category.
- The *non-erotic* category faces do not have any sexual intent or pornographic context, and these are normal expressions such as happy, angry, sad, disgust, neutral, surprise, or fear.

It is worth mentioning that each image in the dataset has been labeled by a single annotator. In some faces, the expressions were both semantically close to erotic and suggestive-erotic. In such confusing samples, we decided on the final category through majority voting by showing the image to two additional annotators.

To sum up, Table 3 depicts how the different stages reduced the initial number of images crawled.

3.3. Limitations in SEA-Faces-30k dataset

We have identified some limitations in the SEA-Faces-30k dataset, which are stated below:

1. The expressions in the dataset are labeled by visual inspection, and they may be improved by Facial Action Coding System (FACS) coding, which is based on facial muscle movement.
2. The erotic expression category contains facial images with oral sexual activities and facial images with only expressions, i.e., without activities.
3. The numbers of male and female faces are not balanced: there are more female faces (i.e., around 80% of the faces are female).

4. Proposed approach

The GAN framework introduced in this paper is an extension of previously proposed GANs: Conditional GAN (CGAN) [56], Semi-Supervised GAN (SSL-GAN) [82,64,15,86], Triple-GAN [46], EC-GAN [34], and BigGAN [8]. Hence, we will first review briefly these network architectures and then introduce our proposed Triple-BigGAN network.

4.1. Preliminaries

4.1.1. GAN and conditional GAN

A basic GAN framework contains two neural networks trained in opposition to one another. Let X denote the real samples and $\mathcal{G}$ denote the generator which takes as input a random noise vector $z \in \mathbb{R}^z$ sampled from a prior noise distribution P_z, uniform or normal, and outputs a synthesized image $\tilde{x} = \mathcal{G}(z) \in \mathbb{R}^d$. Let $\mathcal{D}$ denote the discriminator, which receives an image x as input, which may be either real or synthesized by the generator, and yields a probability distribution, i.e. $\mathcal{D}(x) = P(S|x)$. Ideally, $\mathcal{D}(x) = 1$ when $x \in X$ and $\mathcal{D}(x) = 0$ when x is a synthetic image, i.e. $x = \tilde{x} = \mathcal{G}(z)$. The GAN objective function is given by:

$$\mathbb{E}_{x\sim P_x}[log\mathcal{D}(x)] - \mathbb{E}_{z\sim P_z}[log(1 - \mathcal{D}(\mathcal{G}(z)))], \quad (1)$$

where $\mathbb{E}_x$ represents the expected value over all the data samples. The conditional generative adversarial network [56] is an extension of the GAN in which both $\mathcal{D}$ and $\mathcal{G}$ receive an additional vector of information y as input. The conditional GAN objective is given by:

$$\mathbb{E}_{(x,y)\sim P_{(x,y)}}[log\mathcal{D}(x,y)] - \mathbb{E}_{z\sim P_z}[log(1 - \mathcal{D}(\mathcal{G}(z,y),y)] \quad (2)$$

4.1.2. Semi-supervised GAN and Triple-GAN

A common approach to semi-supervised learning is to combine a supervised and unsupervised objective function during training [82]. As a result, unlabeled data can be leveraged to learn a good representation. In [70], authors have demonstrated that during GANs training, the discriminator learns image representations hierarchically, which may be helpful for object classification. Following this, a simple and useful semi-supervised learning approach can be created by combining unsupervised and supervised GAN objectives.

Let us assume that there are K classes in the labeled data. In most previous works, to extend standard GANs to semi-supervised GAN (i.e. to utilize labeled and unlabeled data), the discriminator output is modified to have K outputs corresponding to real classes [82]. In some works, an additional $(K+1)^{th}$ class corresponding to the fake data generated by the generator is added [74,22], and the discriminator learns by classifying the data among $K+1$ classes.

Despite the success of the technique, it has limitations. For example, the generator does not have much control in deciding the semantics of the generated synthetic samples. Moreover, it may not be possible to have a generator and a discriminator which is also a $(K+1)$-class classifier, both optimal at the same time [61]. The problem appears because when the generator is optimal, it must generate a sample exactly similar to some class among the K *non-fake* classes. At the same time, an optimal discriminator will have two conflicting objectives: to classify this synthetic sample as fake or to classify the same sample among some class among the K *non-fake* classes. Thus, even if the generator was not optimal and the generated sample was similar to some class, the optimal discriminator would still have to contradict objectives to classify it as belonging to some class or as fake. This contradiction justifies that a robust and accurate classifier can not be guaranteed with this kind of generator-discriminator setting.

To overcome these issues, authors in [46], introduced another module along with a conditional generator in GAN called a classifier (i.e. a conditional network). The task of the generator is to generate pseudo samples using the true labels. On the other hand, the classifier has been designed to generate pseudo labels for the true input samples. In the Triple-GAN architecture, the role of the discriminator is only to decide if the sample is real or fake. The classifier performs the task of classifying samples among K classes. To train the discriminator, the authors utilized the labels obtained by the classifier for unlabeled data and also the supervision from the classification loss on the labeled samples. This way, the discriminator is able to guide the generator in an improved way, to generate samples for the respective classes.

4.1.3. BigGAN

When Brock et al. designed BigGAN [8], they adopted a very large-scale generator and discriminator as a class-conditional GAN with a lot of trainable parameters to be able to capture fine details in the synthesized samples. The major focus of the BigGAN was to find a bag-of-tricks based on the best practices in the literature, increasing the batch size (i.e. up to eight times) and the number of parameters (i.e., two to four times). The ultimate goal was to generate realistic high-resolution and high-quality syn-

Table 3
Data Statistics: Number of images on each class of the SEA-Faces30k after every step followed on its creation.

Stage	Num. erotic	Num. suggestive-erotic	Num. non-erotic
(i) Data crawling	31487	30333	18352
(ii) Near duplicate removal	27745	27954	17975
(iii) Faces cropped	22734	22143	15324
(iv) Manual Categorization	10399	10160	10258

thetic images. Through various experiments, the paper demonstrates that the strategies of increasing the batch size and using more model parameters yield better results than the previous state-of-the-art.

As a baseline model, the BigGAN adopted the Self-Attention Generative Adversarial Networks (SAGAN) architecture [90], and it also adopted hinge loss [83,50] as an adversarial loss function, which is similar to SAGAN. Furthermore, the authors adopted the Truncation Trick, originally proposed in [54], and Off-Diagonal Orthogonal Regularization, which is a variant of the Orthogonal Regularization proposed in [9]. To utilize class information in the generator, the BigGAN exploited the class-conditional batch normalization [17], and to utilize it in the discriminator they adopted the projection discriminator [59]. The BigGAN optimization follows the SAGAN guidelines [90] and employs spectral normalization [57], however, different from SAGAN, BigGAN took two steps of discriminator per generator step. To initialize the latent vector z, the orthogonal initialization approach [76] has been used, which has been demonstrated in previous works to be better than the uniform and the Gaussian initialization for Fully Connected (FC) layers. BigGAN utilized a variant of hierarchical latent spaces, and chunks of z are added at multiple layers of the generator as a conditioning vector at different depths to help the generator make better decisions on what to synthesize.

Finally, they show that their proposal can generate high resolution (i.e. 256×256 and 512×512) with high quality too. In summary, the major contributions of the authors are the design strategies for the discriminator and generator network architecture and their training process to finally develop a larger conditional GAN to learn much finer details in the data.

4.2. Proposed Triple-BigGAN

We will first formulate the semi-supervised learning setting adopted in this paper. We denote the images in the dataset as $X = \{x_1, \ldots, x_N\}$, $x_i \in \mathbb{R}^{r \times c}$. Let us assume our dataset contains N images, out of which N_L images contain ground-truth class labels $y_i \in \{1, 2, \ldots, K\}$, whereas N_{UL} images do not contain class labels, i.e., $N = N_L + N_{UL}$. In our SFER problem, just a small portion of the data is labeled, i.e. $N_L > 0$ and $N_L \ll N_{UL}$.

Let the distribution of real samples be $x \sim P_{data}$, the distribution from which the latent vector z is sampled be $z \sim P_z$, the joint distribution of images and their labels be $P(x, y)$, the marginal distribution of images be $P(x)$, the conditional distribution of the class labels given to images be $P(y|x)$ and the conditional distribution of images given by class labels be $P(x|y)$.

In this paper, we propose the Triple-BigGAN framework, an extension of the BigGAN network, for both image classification and generation of class-conditional images with high quality through Semi-Supervised Learning (SSL). The major goal of SSL is to use an easily available large amount of unlabeled data (e.g. faces extracted from pornographic and non-pornographic images unlabeled for facial expressions) to improve the performance of the model for the target problem when the labeled data is not enough to learn representations which are robust and can generalize to unseen samples. The network architecture of the proposed Triple-BigGAN is depicted in Fig. 4. As it is shown, Triple-BigGAN is composed of three parts: a class-conditional generator $\mathcal{G}$, the discriminator $\mathcal{D}$, and a classifier $\mathcal{C}$. The training scheme of our model follows adversarial learning similar to BigGAN. However, we have redesigned the training strategy according to the network modules present in our approach, i.e., discriminator, classifier, and generator.

The goal of the generator $\mathcal{G}$ is to produce synthetic samples, which are conditioned on the class labels in the target data. During the training, we utilize the complete dataset i.e., labeled as well as unlabeled data (P_{data}) to learn $G(z, y)$, which can generate samples similar to $P(x|y)$, and for this, we provide as input a latent vector $z \sim P_z$ and a class label $y \sim P(y)$. The output generated by $\mathcal{G}$ can be considered as $x|y \sim P_g(x|y)$ for some given $y \sim P(y)$. We can consider this pseudo input-label pair output as $(x_g, y_g) \sim P_g(x, y)$.

The inputs to the classifier $\mathcal{C}$ are both labeled and unlabeled data. The labeled data will be used to provide supervision for the classifier, whereas the unlabeled data will be used to draw $(x_c, y_c) \sim P_c(x, y)$, which can be considered as pseudo input-label pairs.

Finally, the job of the discriminator $\mathcal{D}$ is to differentiate the real image-label pairs $(x_l, y_l) \in P_{data}(x, y)$ from the fake sample-label pairs obtained by the generator's fake samples i.e., $\mathcal{G}$ i.e., $(x_g, y_g) \sim P_g(x, y)$ or the labels estimated by the classifier for the unlabeled input images i.e., $(x_c, y_c) \sim P_c(x, y)$.

The overall objective of the combined network modules is to learn a classifier that can output labels for the data accurate enough to consider them equivalent to the ground truth labels, i.e. $P_c(x, y) \sim P_{data}$, and to learn a generator that can generate synthetic call-conditional samples similar to the true data distribution, i.e. $P_g(x, y) \sim P_{data}$. The whole network attains convergence when the objects of the classifier and the generator are achieved successfully. In the proposed architecture, the labels predicted by the classifier for the unlabeled images help the generator to learn a class conditional representation similar to true data distribution. Similarly, the high-fidelity samples synthesized by the generator help the classifier to yield better classification performance on the unlabeled data. Therefore, the proposed Triple-BigGAN model is able to improve both instance synthesis and classification in the semi-supervised setting.

Concretely, the discriminator loss $\mathcal{L}_{\mathcal{D}}$, the generator loss $\mathcal{L}_{\mathcal{G}}$, and the classifier loss $\mathcal{L}_{\mathcal{C}}$ are defined as follows:

$$\begin{aligned}\mathcal{L}_{\mathcal{D}} = {} & \mathbb{E}_{(x,y)\sim P_{data}}[min(0, -1 + \mathcal{D}(x, y))] - (1 - \gamma) \\ & \cdot \mathbb{E}_{z\sim P_z, y\sim P_{data}}[min(0, -1 - \mathcal{D}(\mathcal{G}(z, y), y))] - \gamma \\ & \cdot \mathbb{E}_{x\sim P_{data}}[min(0, -1 - \mathcal{D}(x, \mathcal{C}(x))]\end{aligned} \tag{3}$$

$$\mathcal{L}_{\mathcal{G}} = -\mathbb{E}_{z\sim P_z, y\sim P_{data}}\mathcal{D}(\mathcal{G}(z, y), y)) \tag{4}$$

$$\begin{aligned}\mathcal{L}_{\mathcal{C}} = {} & \mathbb{E}_{(x,y)\sim P_{data}}[min(0, -1 + \mathcal{C}(x, y))] - \mathbb{E}_{z\sim P_z, y\sim P_{data}}[min(0, -1 \\ & - \mathcal{C}(\mathcal{G}(z, y), y))],\end{aligned} \tag{5}$$

where γ is a parameter that assigns relative weights to the generator and classifier. In our experiments, we assigned the same weights to the classifier and generator.

4.3. How Triple-BigGAN differs from other GANs

Some recent works have introduced inference networks in GANs. For instance, InfoGAN [12] learns explainable latent codes from unlabeled data by regularizing the original GANs via variational mutual information maximization. Dumoulin et al. [22] presented the Adversarially Learned Inference (ALI) model. In this model, the inference network approximates the posterior distribution of latent variables given true data in an unsupervised manner. The discriminator will reject the samples from the classifier in Triple-BigGAN while the discriminator will accept the samples from the inference network in ALI, which leads to different update rules for the discriminator and inference network.

Our work extends the Triple-GAN, EC-GAN, and BigGAN frameworks, while showing significant differences with them. Triple-BigGAN follows the adversarial training methodology of Triple-GAN, but we have redesigned the classifier, generator, discriminator, and loss functions to generate high-fidelity class-conditional data distributions and an improved classifier. The principal objec-

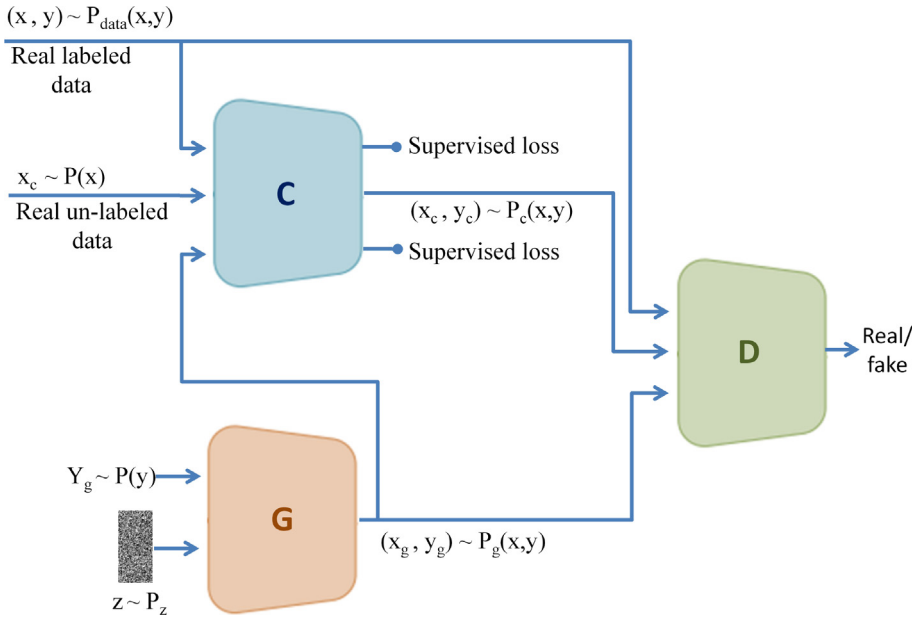

Fig. 4. Illustration of the proposed Triple-BigGAN. Triple-BigGAN has a generator, a discriminator, and a classifier. The classifier is trained on the image-label pairs in real labeled data set as well as generated by the generator. The discriminator's job is to detect image-label pairs in real labeled data set as real and image-label pairs obtained from the classifier and generator for the unlabeled dataset as fake.

tive of the Triple-GAN approach is to train a robust discriminator, and the EC-GAN aims mainly to train an improved classifier exploiting the GAN part. Differently from Triple-GAN and EC-GAN, TripleBig-GAN focuses on training both a robust classifier and discriminator in an end-to-end manner.

Furthermore, our work differs from BigGAN as per details provided in Sections 4.1.3 and 4.2. Briefly, BigGAN focuses on unsupervised image synthesis and has only a discriminator and a generator, whereas Triple-BigGAN aims at semi-supervised joint distribution matching. Our network can utilize labeled and unlabeled data and learns image classification and image synthesis, both together in an end-to-end manner through a discriminator, a generator, and a classifier.

5. Experiments and results

5.1. Experimental setup

To verify the proposed model, first, we empirically verified the image synthesis capabilities of Triple-BigGAN using the SEA-Faces-30k dataset (see Section 3) to investigate the quality of the synthesized human face images. Then, we evaluated the performance of Triple-BigGAN in semi-supervised image classification on the FER task, using the SEA-Faces-30k, Expression in-the-Wild (ExpW) [91], and FER-2013 datasets. Finally, we evaluated the classification performance on three general-purpose image classification benchmark datasets, i.e., Modified National Institute of Standards and Technology dataset (MNIST) [45], CIFAR-10 [43], and Street View House Numbers (SVHN) [62]. Furthermore, we also performed an additional experiment to evaluate the usefulness of synthetic images generated by Triple-BigGAN to improve the accuracy of deep CNN networks by augmenting the training data.

The ExpW dataset contains 91,795 facial images manually labeled with seven expression categories: angry, disgust, fear, happy, sad, surprise, and neutral. The images in the dataset have been collected from the Google searcher. The dataset does not provide predefined training and testing sets.

FER-2013 is also a standard FER research dataset collected through Google Search API. The dataset contains 35,887 face images with 28,709 training, 3,589 validation, and 3,589 test images. The images have been re-scaled in the dataset to 48×48 pixels after adjusting the cropped area. Like ExpW, FER-2013 also has seven expressions: anger, disgust, fear, happiness, sadness, surprise, and neutral.

MNIST dataset has 50,000 and 10,000 images for training and validation, respectively. There are another 10,000 images for test purposes. The images in the dataset are handwritten digits of 28×28 pixels resolution. The SVHN dataset contains 73,257 training and 26,032 test images, respectively. Each one is an RGB sample with a resolution of 32×32 of numbers with varying backgrounds. In the CIFAR-10 dataset, there are RGB images from 10 different classes: automobile, aeroplane, bird, cat, deer, dog, frog, horse, ship, and truck. It consists of 50,000 training and 10,000 test images, each one with a resolution of 32×32. Since a separate validation set is not given in SVHN and CIFAR-10 datasets, it can be extracted from the training set, if required.

In the case of the MNIST, SVHN, and CIFAR-10 datasets, we adopted the same settings that have been adopted by many previous works [86,46,85,22,82,15]. Specifically, we performed experiments for the cases in which there are 100, 1000, and 4000 randomly selected labeled test instances, respectively. In each of these cases, random sampling has been carried out ten times, and we reported the mean and standard deviation of the test error rates for the classification task. Moreover, we have compared Triple-BigGAN with several methods by taking their results on these three datasets from the existing literature. These methods are: EnhancedTGAN [86], Triple-GAN [46], CT-GAN [85], ALI [22], CatGAN [82], and GoodBadGAN [15].

In the case of the novel SEA-Faces-30k dataset, for a fair comparison of the SFER classification performance, we trained models using four inference-based GANs, i.e., Triple-GAN, CatGAN, ALI, and GoodBadGAN. First, we divided the dataset into train, validation, and test sets by randomly taking 70%, 15%, and 15% images, respectively. Then, the results for inference GANs are calculated using two different sizes of training datasets: (i) randomly selecting 3000 labeled images from the training set to understand its

capabilities with a lesser amount of training data and (ii) using the complete train set.

Apart from the GAN-based methods, we also fine-tuned a state-of-the-art face recognition approach, i.e. FaceNet [77], and two famous deep CNNs, i.e., VGG-16 [80] and ResNet-50 [36]. FaceNet is based on the Inception-ResNet v1 network trained with triplet loss, and we utilized the publicly available weights (i.e., the network pre-trained using face datasets). In the case of the VGG-16 and ResNet-50 networks, we utilized the models pre-trained with the ImageNet dataset, which we further fine-tuned with a large-scale face dataset, i.e., CASIA-WebFace[5], for the face classification task. Then, for the SFER task, we first extracted features from the average pooling layer in Inception-ResNet v1 and ResNet-50 networks, and, in the case of VGG-16, we extracted the features from the last max-pooling layer. Next, using the extracted features as input, we trained a Multi-Layer Perceptron (MLP) for each of the above three networks, using the train and validation sets of SEA-Faces-30k. The MLP network is designed with two residual blocks with skip connections, i.e., four layers in total.

Furthermore, we also evaluated a recent Deep Learning-based approach: CovPoolFER [1], a model specifically designed for facial expression recognition[6]. In CovPoolFER, the authors initially extracted and flattened deep features, and then they summarized the second-order information in the feature set through the computation of a covariance matrix. Finally, they fed the encoded features to a Symmetric Positive Definite (SPD) Manifold Network (SPDNet) layer for dimensionality reduction and non-linearity on covariance matrices. During the evaluation, we fine-tuned CovPoolFER on the SEA-Faces-30k train set.

In addition, we evaluated the Triple-BigGAN on the SEA-Faces-30k dataset to investigate the quality of the synthesized human face images. For the image synthesis task, we did a comparison of our approach with three recent state-of-the-art approaches, i.e., Triple-GAN, BigGAN, and StyleGAN. The evaluation is performed using the Inception Score (IS) [74] and the Frechet Inception Distance (FID) [37]. In the case of the IS, the higher it is, the better the synthetic image is considered, whereas the lower the FID, the better the synthetic image.

All the experiments have been carried out using Python 3.6, Keras 2.3.0, PyTorch 1.5, and TensorFlow 1.14, with four Tesla K40 (12 GB) and two Tesla K80 GPUs (24 GB).

5.2. Implementation and network training

The generator and discriminator in Triple-BigGAN closely follow the network structures in BigGAN, and their architecture details, adopted in the experimental analysis of this paper, are given in Tables 4 and 5, respectively. Tables 6 and 7 show the architecture details of the Residual blocks used in Triple-BigGAN, i.e., ResBlock up in generator and ResBlock down in discriminator, respectively. In the Tables, H and W represent the height and width of the input, and C_{in} and C_{out} are the number of input and output channels. The ResBlock in the last layer of the discriminator (i.e., without downsampling) does not contain the skip connection layer. In the classifier, the network adopted is ResNet-50 [36] and at its global average pooling output layer, an MLP network is attached which is created using two residual blocks with skip connections, i.e., four layers in total.

The training of our Triple-BigGAN network follows the guidelines for the SAGAN [90] and BigGAN [8] networks. The weights in Triple-BigGAN's generator, discriminator, and classifier networks are initialized through orthogonal initialization [76]. We used the Adam [42] optimizer (momentum parameters $\beta_1 = 0, \beta_2 = 0.999$) with the learning rate $5 \cdot 10^{-5}$ for the generator, $2 \cdot 10^{-4}$ for the discriminator, and $2 \cdot 10^{-4}$ for the classifier. We also utilized spectral normalization [57] and Orthogonal Regularization [9] in the generator and discriminator (but not in the classifier) for training stability. In the discriminator, spectral normalization is used in all weight layers. The latent vector z (i.e., the random noise as input to the generator) is drawn from the normal distribution $\mathcal{N}(0, I)$.

Table 4
Architecture for Triple-BigGAN's Generator. Note that "*ch*" is the channel width multiplier (i.e. $ch = 128$ in Triple-BigGAN). "BN" stands for batch normalization and "SN" denotes Spectral Normalization.

Layer/Block	SN	#output
$z \in \mathbb{R}^{128} \sim \mathcal{N}(0, I)$	-	128
Embed(y) $\in \mathbb{R}^{128}$	-	128
Dense $(128 + 128) \rightarrow 16 \cdot ch$	-	$4 \times 4 \times 16 \cdot ch$
ResBlock up $16 \cdot ch \rightarrow 16 \cdot ch$	Y	$8 \times 8 \times 16 \cdot ch$
ResBlock up $16 \cdot ch \rightarrow 8 \cdot ch$	Y	$16 \times 16 \times 8 \cdot ch$
ResBlock up $8 \cdot ch \rightarrow 8 \cdot ch$	Y	$32 \times 32 \times 8 \cdot ch$
ResBlock up $8 \cdot ch \rightarrow 4 \cdot ch$	Y	$64 \times 64 \times 4 \cdot ch$
ResBlock up $4 \cdot ch \rightarrow 2 \cdot ch$	Y	$128 \times 128 \times 2 \cdot ch$
Non-Local Block		$128 \times 128 \times 2 \cdot ch$
ResBlock up $2 \cdot ch \rightarrow 1 \cdot ch$	Y	$256 \times 256 \times 1 \cdot ch$
BN, ReLU, 3×3Conv $ch \rightarrow 3$	–	$256 \times 256 \times 3$
Tanh	–	$256 \times 256 \times 3$

Table 5
Architecture for Triple-BigGAN's Discriminator network. Note that "*ch*" is the channel width multiplier (i.e. $ch = 128$ in Triple-BigGAN). "y" stands for the class labels, and "h" denotes the previous layer's output.

Layer/Block	#output
Input image	$256 \times 256 \times 3$
ResBlock down $3 \rightarrow ch$	$128 \times 128 \times 1 \cdot ch$
ResBlock down $ch \rightarrow ch$	$64 \times 64 \times 1 \cdot ch$
Non-Local Block	$64 \times 64 \times 1 \cdot ch$
ResBlock down $ch \rightarrow 2 \cdot ch$	$32 \times 32 \times 2 \cdot ch$
ResBlock down $2 \cdot ch \rightarrow 4 \cdot ch$	$16 \times 16 \times 4 \cdot ch$
ResBlock down $4 \cdot ch \rightarrow 8 \cdot ch$	$8 \times 8 \times 8 \cdot ch$
ResBlock down $8 \cdot ch \rightarrow 16 \cdot ch$	$4 \times 4 \times 16 \cdot ch$
ResBlock $16 \cdot ch \rightarrow 16 \cdot ch$	$4 \times 4 \times 16 \cdot ch$
ReLU, Global sum pooling	$1 \times 1 \times 16 \cdot ch$
Embed(y)$\cdot h + (dense \rightarrow 1)$	1

Table 6
Architecture of the Residual Block in Triple-BigGAN's Generator (i.e., ResBlock up in Table 4).

Layer	Kernel	#output
shortcut/skip	[1, 1, 1]	$2 \cdot H \times 2 \cdot W \times C_{out}$
BN, ReLU	–	$H \times W \times C_{in}$
Conv	[3, 3, 1]	$2 \cdot H \times 2 \cdot W \times C_{out}$
BN, ReLU	–	$2 \cdot H \times 2 \cdot W \times C_{out}$
Conv	[3, 3, 1]	$2 \cdot H \times 2 \cdot W \times C_{out}$
Addition	–	$2 \cdot H \times 2 \cdot W \times C_{out}$

Table 7
Architecture of Residual Block in Triple-BigGAN's Discriminator (i.e., ResBlock down in Table 5).

Layer	Kernel	#output
shortcut/skip	[1, 1, 1]	$H/2 \times W/2 \times C_{out}$
ReLU	–	$H \times W \times C_{in}$
Conv	[3, 3, 1]	$H \times W \times C_{out}$
ReLU	–	$H \times W \times C_{out}$
Conv	[3, 3, 1]	$H/2 \times W/2 \times C_{out}$
Addition	–	$H/2 \times W/2 \times C_{out}$

[5] http://www.cbsr.ia.ac.cn/english/CASIA-WebFace-Database.html
[6] https://github.com/d-acharya/CovPoolFER

As shown in Table 4, the $z \in \mathbb{R}^{128}$ is concatenated with the class embedding. Then the combined vector is passed in residual blocks via the skip connections.

During training, we performed two discriminator steps for each generator and classifier step, and training is done for 500k steps with a batch size of 256. Due to memory constraints, we could not try larger batch sizes. The z is concatenated with the class label embedding, and the output vector is sent to the residual blocks via skip connections.

To overcome overfitting and to have more training samples, we performed data augmentation with class-preserving transformations. First, for data enlargement, we cropped facial regions with four different margins around the detected face bounding box: 20, 40, 60, and 80 pixels. After such crops, the faces are re-scaled to 320×320 pixels, resulting in four scales of the input images. Then, we performed horizontal and vertical translations by 20% and horizontal flip. All the images are randomly rotated in the range of $\pm 30^{\circ}$. Therefore, the number of images obtained is 4 (scales) $\times$ 4 (translations) $\times$ 2 (flip) = 32 times the original images. Thereafter, the images were resized according to the input size of the networks, i.e., 256×256.

5.3. Triple-BigGAN for Synthetic Image Generation: SEA-Faces-30k Dataset

First, we evaluated our model for synthetic image generation. It has been trained using the SEA-Faces-30k dataset, employing the settings mentioned in 5.2.

Some samples generated by Triple-BigGAN for the "erotic" and "suggestive-erotic" classes are shown in Figs. 5 and 6, respectively.

It can be noticed that our model can synthesize images with a high-quality and large variety of content. We also report a comparative analysis of the FID and IS values obtained by our approach and by four recent GAN networks in Table 8, i.e., Triple-GAN [46], SAGAN [90], BigGAN [8], and StyleGAN [41]. It can be noticed that our proposed approach either outperformed or provided results comparable to the other state-of-the-art models in terms of IS and FID scores.

Furthermore, we also found that all the models generated some wrong images. Some examples of badly generated images by Triple-BigGAN are shown in Fig. 7. We observed that the significant mistakes in the generated samples were local, i.e., mainly artifacts, or images consisting of texture blobs instead of objects. The generation of the suggestive-erotic class images is more challenging because of not having features as intense as in the erotic class and some similarity/overlapping with both the erotic and the non-erotic categories. We also found that the issues in the synthetic images were bigger when the complexity of the images in the training set was high, e.g., there were faces with occlusions, highly posed, or with low resolution.

5.4. Triple-BigGAN for facial expression recognition: SEA-Faces-30k Dataset

The pipeline for classification of the expressions in facial images comprises the following stages: (i) face detection, (ii) facial region representation, and (iii) classification of the encoded data in three categories: erotic, suggestive-erotic, or non-erotic. For the *face detection*, we used the RetinaFace detector, as we did during the SEA-Faces-30k generation (see Section 3).

In this experiment, SEA-Faces-30k has been randomly divided into training, test, and validation sets, which have 70%, 15%, and 15% of the images of the dataset, respectively.

For comparative analysis, we utilized three handcrafted descriptor-based approaches, i.e., LBP [78], HOG [66], SIFT [52], two Deep CNN networks, i.e., VGG-16 [80] and ResNet-50 [36], a deep learning-based FER approach, i.e., CovPoolFER [1], and four inference based SSL GANs, i.e., ALI [22], CatGAN [82], Triple-GAN [46] and GoodBadGAN [15]. Following the previously adopted protocols in [46,22,74], we performed the evaluation under two settings: a) with 3000 labeled images (i.e., 1000 from each category) as a labelled set and left out samples in the training set as an unlabeled set and, b) when using all the labeled images in SEA-Faces-30k dataset.

Table 9 presents the error rate in the classification of sexual expressions achieved by Triple-BigGAN and the other assessed approaches. We can observe that Triple-BigGAN obtains the lowest error rate under both evaluation settings, i.e., 15.77% when only 3000 labeled examples were used, and 6.42% when all the labeled samples were used. It empirically justifies the better learning capabilities of the proposed network. We can also notice that deep features consistently outperform the results obtained by the traditional image descriptors. It is also remarkable that the methods that utilized unlabeled data along with labeled data. i.e. ALI, Triple-GAN, and Triple-BigGAN provided better performance than the other approaches. We attribute this performance improvement to the learning of a better representation because of the capabilities of these methods to exploit information from the unlabeled data in addition to the labeled data. Amongst the non-deep features-based methods, HOG combined with MLP achieved the best error rates, i.e. 45.51% and 28.53%.

Concerning the features obtained by CovPoolFER, the error rates obtained when combining it with MLP, i.e. 28.85% and 11.95%, outperform the results obtained by traditional descriptors, but are lower than the assessed GANs.

It should also be noted that, on average, the deeper CNN architectures provided an improvement in the performance by up to 100% compared to local image descriptors: the best error rates obtained by CNN features are 27.36% and 11.65% against the best descriptor errors of 45.51% and 28.53%.

Additionally, we show the confusion matrices in Fig. 8 for the Triple-BigGAN approach on the SEA-Faces-30k dataset, when only 3000 labels are used and when all the labeled data is used. In both cases, the highest number of misclassifications has happened between the classes "erotic" and "suggestive-erotic" and between "non-erotic" and "suggestive-erotic". These results show that the "suggestive-erotic" category is more difficult to detect than the non-erotic and erotic categories, mainly because of the overlap of this class with the other two.

5.5. Triple-BigGAN generated labeled Samples for SFER training data augmentation

This analysis aims to study the impact of data augmentation – through adding synthesized images to the existing dataset – on the SFER accuracy. We generated 5,000 synthetic images using Triple-BigGAN for each category and augmented the SEA-Faces-30k dataset to 45k images. The results of this assessment are presented in Table 10. It is illustrated that additional data generated by GAN helps to increase the performance of deep CNNs by more than 3 percentage points. These results show that Triple-BigGAN can generate images with good quality and different variations not seen in the training dataset.

5.6. Triple-BigGAN for facial expression recognition: benchmark datasets

Through the experiments in this section, we evaluated the proposed Triple-BigGAN network on two standard benchmark datasets for Facial Expression Recognition: FER-2013 and ExpW datasets. For comparative analysis of Triple-BigGAN, we utilized

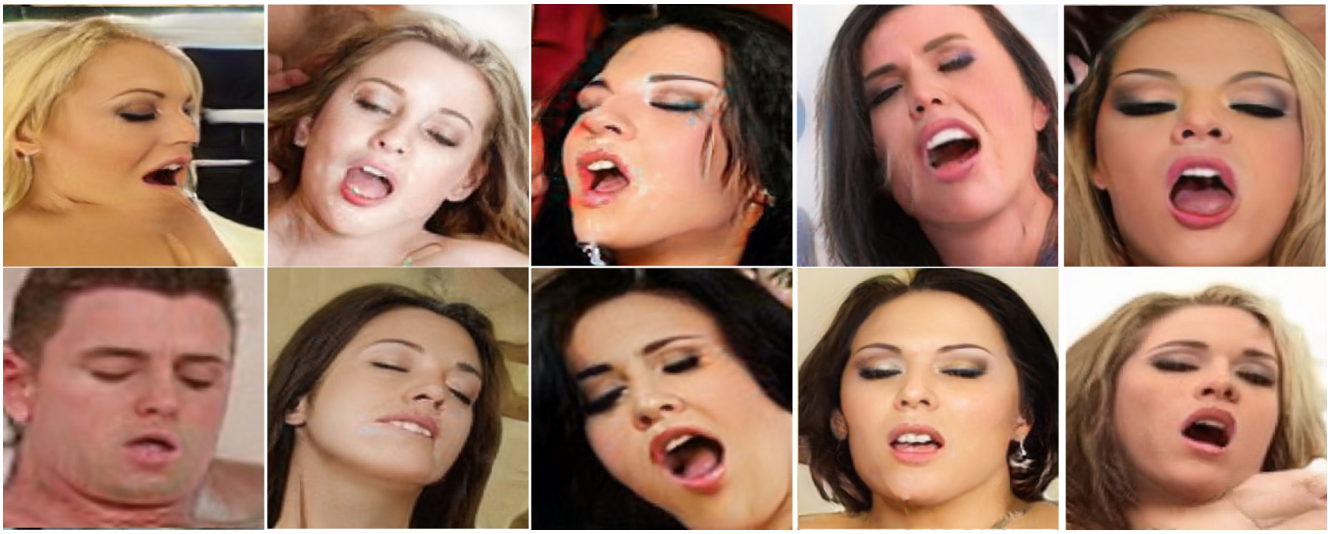

Fig. 5. Samples of images from the class "erotic" generated By Triple-BigGAN. Their resolution was 256 × 256 pixels.

Fig. 6. Samples of images from the class "suggestive-erotic" generated by Triple-BigGAN. Their resolution was 256 × 256 pixels.

Table 8
Comparative Analysis of Frechet Inception Distance (FID) and Inception Score (IS) scores on SEA-Faces-30k dataset.

Approach	Resolution	FID	IS
SAGAN	128 × 128	38.56	52.78
Triple-GAN	64 × 64	42.04	34.89
BigGAN	256 × 256	19.97	97.68
StyleGAN	256 × 256	21.83	89.32
Triple-BigGAN	256 × 256	**19.94**	**97.98**

VGG-16 [80], ResNet-50 [36], CovPoolFER [1], ALI [22], and Triple-GAN [46]. Moreover, some results on these datasets have also been taken for the recent state-of-the-art approaches directly from the respective published papers.

The FER-2013 dataset consists of 35,887 face images with 28,709 training, 3,589 validation, and 3,589 test images. In the case of the ExpW dataset, as the train, validation, and test sets are not provided explicitly, similar to [4], we divided it into the train, validation, and test sets by randomly taking 80%, 10%, and 10% images, respectively. Then, the semi-supervised training of the networks was performed under two settings: (i) randomly selecting 3000 labeled images from the training set to understand its capabilities with a lesser amount of training data and (ii) using the complete train set. To select 3000 samples from the training set, we utilized the ground-truth confidence score given for the faces in the dataset, and we selected faces with confidence greater than 60.

The results for Triple-BigGAN on FER-2013 and ExpW datasets are reported in Table 11. From the results, it can be seen that an excellent accuracy is reported by the Triple-BigGAN, especially on the FER-2013 dataset, the Triple-BigGAN has obtained the best results in both the settings i.e., with "3000 labels" and "all the labels". On the ExpW dataset, our model achieves the best accuracy compared to all the approaches evaluated and many state-of-the-art methods from the literature except THIN [2]. As there is no standard test set in ExpW dataset for evaluation, the result of the THIN method has been reported with 6673 samples in the test set compared to a much larger test set in our work containing 9179 samples. On both datasets, in the case of 3000 labeled samples, our network has obtained much superior accuracies com-

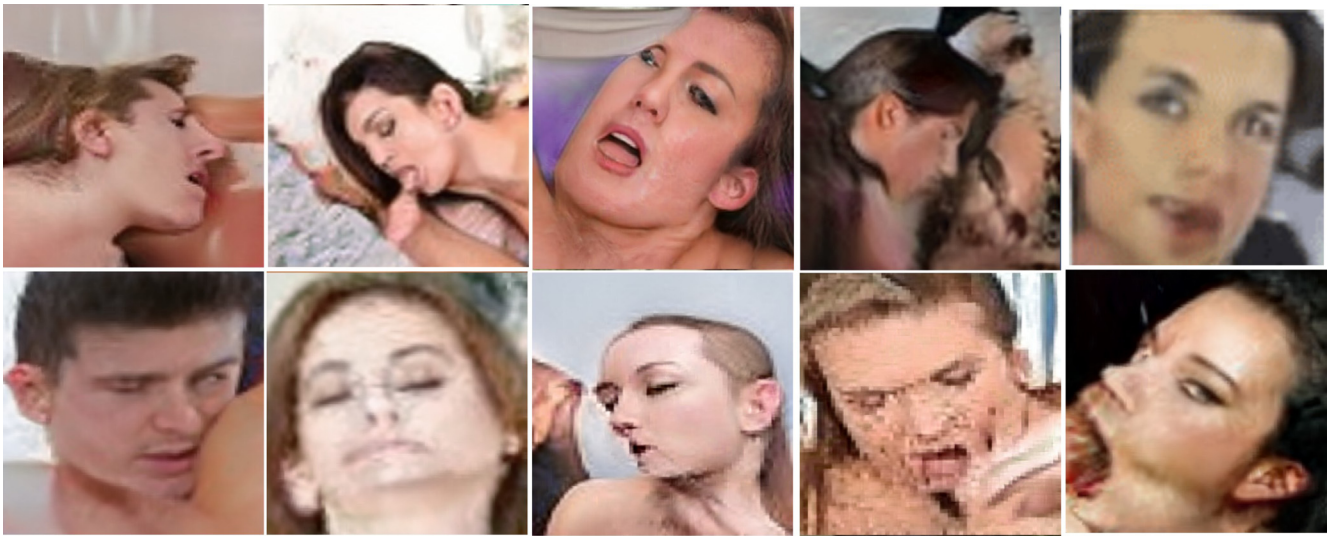

Fig. 7. Samples of wrong images generated by Triple-BigGAN. Their resolution was 256 × 256 pixels.

Table 9
Comparative analysis of Triple-BigGAN for Sexual Facial Expression Recognition. ULD refers to whether the method utilizes unlabeled data (Y) or not (N).

Approach	Test error rate (%) with # labels		
	ULD	3000 labels	All labels
LBP + MLP	N	58.23 ± 5.54	40.10 ± 3.88
HOG + MLP	N	45.51 ± 4.89	28.53 ± 2.87
SIFT + MLP	N	62.30 ± 5.67	45.30 ± 3.34
VGG-16 + MLP	N	27.36 ± 2.39	11.65 ± 1.20
ResNet-50 + MLP	N	29.71 ± 2.56	13.10 ± 1.28
CovPoolFER + MLP	N	28.85 ± 2.89	11.95 ± 1.32
CatGAN	N	25.24 ± 2.34	17.88 ± 1.22
GoodBadGAN	N	26.21 ± 2.40	20.76 ± 2.03
ALI	Y	23.73 ± 2.83	15.79 ± 1.12
Triple-GAN	Y	19.09 ± 1.98	11.98 ± 0.99
Triple-BigGAN	Y	**15.77 ± 1.19**	**6.41 ± 0.90**

Table 10
SFER Accuracy analysis on SEA-Faces-30k data augmented with synthesized samples.

Approach	Test Error Rate (%)	
	without synthetic images	with synthetic images
VGG-16 + MLP	11.65 ± 1.20	8.25 ± 0.96
ResNet-50 + MLP	13.10 ± 1.28	10.11 ± 1.10
CovPoolFER	11.95 ± 1.33	8.95 ± 1.02

pared to their other counterparts. Also, Triple-BigGAN in this case has reported accuracies almost similar to the accuracies obtained by VGG-16 with all the labeled samples.

Figs. 9 and 10, depict the confusion matrices achieved on the FER-2013 and ExpW datasets, respectively. We can see that when 3000 labels were used as labelled samples (see subfigure B), the "happy" class obtained the highest accuracy and the "fear" category is more challenging to differentiate from the other facial expressions. In the FER-2013 dataset, the "fear" class samples obtained the highest confusion with the "sadness" class and then with the "anger" class. In the ExpW dataset, the "fear" class found the highest closeness with "anger" samples and then next with the "surprise" class.

5.7. Triple-BigGAN for image classification: benchmark datasets

Additionally, we compared Triple-BigGAN with state-of-the-art semi-supervised deep learning models on the MNIST, SVHN, and CIFAR-10 datasets, which are widely used for the evaluation of classification. Following the evaluation settings widely adopted [46,86,85,22,82,15] on these datasets, i.e., 100, 1000, and 4000 labels respectively and also all labels, we used the same methodology for the evaluation of Triple-BigGAN too. The error rates of the competing methods have been taken from the existing literature, except for the Triple-GAN model, for which we did compute the

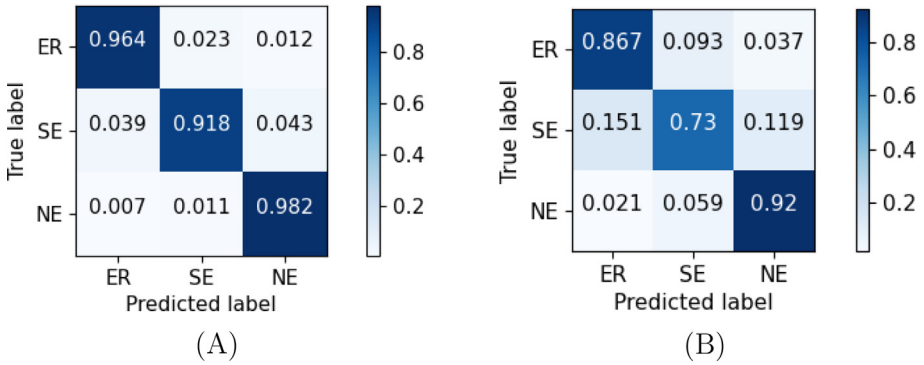

Fig. 8. Confusion matrices for Triple-BigGAN on SEA-Faces-30k: (A) when all labels are used, (B) when 3000 labels are used as labeled sample and remaining as unlabeled. Here, ER represents erotic, SE means suggestive-erotic, and NE is a non-erotic class.

Table 11
Comparative analysis of Triple-BigGAN and the competing methods on the FER-2013 and ExpW datasets. ULD refers to whether the method utilizes unlabeled data (Y) or not (N). The results taken from the literature are shown with a "*" sign and "-" sign indicates that the result is not available.

Approach	ULD	Test error rate (%) with # labels				
		FER-2013		ExpW		
		3000 labels	all labels	#test samples	3000 labels	all labels
VGG-16 + MLP	N	54.01	66.35	9179	52.06	65.01
ResNet-50 + MLP	N	57.49	68.52	9179	53.91	66.16
CovPoolFER + MLP	N	55.34	67.48	9179	52.45	65.73
ALI	Y	60.92	66.08	9179	58.45	66.07
Triple-GAN	Y	61.35	68.89	9179	60.78	67.26
Triple-BigGAN	Y	**65.28**	**73.12**	9179	**64.98**	73.27
ResNet-50(Ad-Corre)[27]*	N	-	72.03	-	-	-
Shao et al.[79]*	N	-	71.14	-	-	-
BReG-NeXt[35]*	N	-	71.53	-	-	-
THIN[2]*	N	-	-	6673	-	**76.08**
EAFR[49]*	N	-	-	6673	-	71.90
PAT-ResNet-101[11]*	N	-	73.28	1400	-	72.93
Benamara et al.[4]*	N	-	72.72	9179	-	71.82

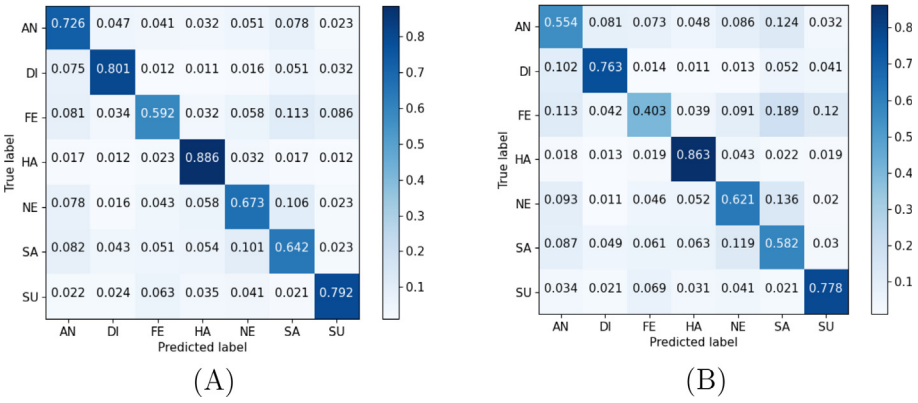

Fig. 9. Confusion matrices for Triple-BigGAN on FER-2013 dataset: (A) when all labels are used, and (B) when 3000 labels are used as labeled sample and remaining as unlabeled. The names in the rows/columns stand for the classes of the dataset: Anger (AN), Disgust (DI), Fear (FE), Happy (HA), Neutral (NE), Sad (SA), and Surprise (SU).

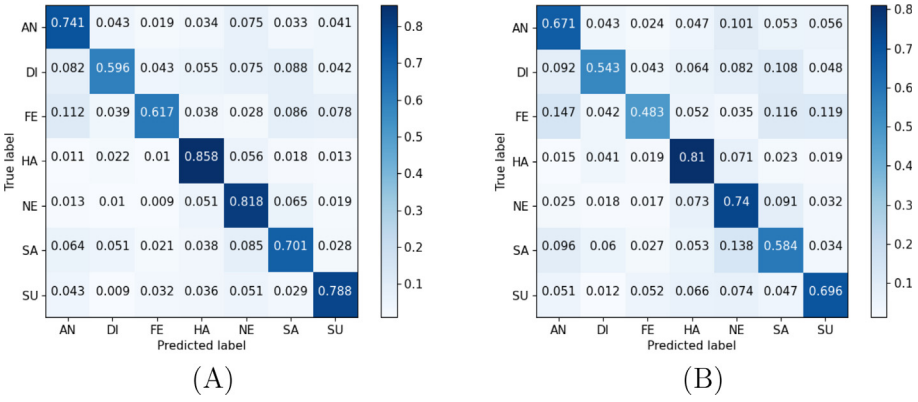

Fig. 10. Confusion matrices for Triple-BigGAN on ExpW dataset: (A) when all labels are used, (B) when 3000 labels are used as labeled sample and remaining as unlabeled. The names in the rows/columns stand for the classes of the dataset: Anger (AN), Disgust (DI), Fear (FE), Happy (HA), Neutral (NE), Sad (SA), and Surprise (SU).

results. The error rates of this classification experiment are presented in Table 12. From Table 12, it is evident that the proposed Triple-BigGAN obtains better performance than all the other approaches. For CIFAR-10, when we used 4000 labels, Triple-BigGAN has shown a significant improvement compared to Triple-GAN, i.e., the test error rate decreased from 16.99% to 8.90%. It is also clearly demonstrated that the proposed Triple-BigGAN obtains better or comparable accuracy in comparison to all the other state-of-the-art methods under all different settings on the three datasets evaluated in this experiment.

Table 12
Comparative analysis of Triple-BigGAN and the competing methods on MNIST, CIFAR and SVHN datasets

Approach	Test error rate (%) with # labels					
	MNIST		SVHN		CIFAR-10	
	100 labels	All labels	1000 labels	All labels	4000 labels	All labels
CatGAN	1.39 ± 0.28	-	-	-	19.58 ± 0.58	-
Improved-GAN	0.93 ± 0.07	-	8.11 ± 1.30		18.63 ± 2.32	-
ALI	-	-	7.42 ± 0.65	-	17.99 ± 1.62	-
Triple-GAN	0.91 ± 0.58	-	5.77 ± 0.17	-	16.99 ± 0.36	-
GoodBadGAN	0.80 ± 0.10	-	4.25 ± 0.03	-	14.41 ± 0.03	-
CT-GAN	0.89 ± 0.13	-	-	-	9.98 ± 0.21	-
EnhancedTGAN	0.42 ± 0.03	0.27 ± 0.03	2.97 ± 0.09	2.23 ± 0.01	9.42 ± 0.22	4.80 ± 0.07
Triple-BigGAN	**0.39** ± 0.029	**0.26** ± 0.02	**2.85** ± 0.07	**2.12** ± 0.01	**8.90** ± 0.21	**4.12** ± 0.06

6. Conclusion and future work

In this work, we have proposed the Triple-BigGAN model to improve both semi-supervised conditional image synthesis and classification. First, we investigated if facial information can be utilized for the Sexual Facial Expression Recognition (SFER) task. Since there was no dataset publicly available for SFER, we introduced a new dataset named as SEA-Faces-30k, which contains challenging images under three categories: erotic, suggestive-erotic, and non-erotic. Then, through a series of experiments, we demonstrated that the proposed framework generates high-quality and high-resolution synthetic images, and also that the synthetic images generated by our approach can improve the error rates for supervised learning-based methods.

To evaluate the quality of the synthetic images generated by Triple-BigGAN, we used the FID and IS scores. Using these scores, we compared Triple-BigGAN with other state-of-the-art competitive approaches, resulting that the Triple-BigGAN network provided comparable or better results than these methods. Then we evaluated the strength of Triple-BigGAN for the novel SFER task using our newly proposed SEA-Faces-30k dataset and for this comparative analysis, we used classical feature extractors as well as modern CNN and GAN-based approaches. Our approach not only obtained a remarkable accuracy of 93.59% for the sexual expression detection task but also outperformed other methods. This justifies empirically that facial information can be exploited for SFER with high accuracy. To the best of our knowledge, this is the first study to detect automatically sexual facial expressions. Additionally, we also compared the classification performance of Triple-BigGAN against inference-based GANs on three image classification benchmark datasets, i.e., MNIST, CIFAR-10, and SVHN. The Triple-BigGAN improved the state-of-the-art results and obtained the best results on all three datasets.

Further, we also assessed how well the Triple-BigGAN models could perform on standard Facial Expression Recognition (FER) benchmark datasets. Triple-BigGAN obtained state-of-the-art results with 73.27% and 73.12% accuracies on ExpW and FER-2013 datasets, respectively.

In future works, we will extend the SEA-Faces-30k dataset and improve its limitations, We will also validate the sexual expression recognition methods on Child Sexual Exploitation Material (CSEM), and assess if SFER models are useful for boosting existing pornography and, subsequently, CSEM detection methods.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research has been funded with support from the European Union's Horizon 2020 Research and Innovation Framework Programme, H2020 SU-FCT-2019 under the GRACE project with Grant Agreement 883341. This publication reflects the views only of the authors, and the European Union's Horizon 2020 Research and Innovation Framework Programme, H2020 SU-FCT-2019 cannot be held responsible for any use which may be made of the information contained therein

References

[1] D. Acharya, Z. Huang, D.P. Paudel, L. Van Gool, Covariance pooling for facial expression recognition, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, volume 2018-June, pages 480–487.
[2] E. Arnaud, A. Dapogny, K. Bailly, Thin: Throwable information networks and application for facial expression recognition in the wild, IEEE Trans. Affect. Comput. (2022) 1.
[3] R. Basson, Human sexual response, Handbook of Clinical Neurology, volume 130, Little, Brown, 2015, pp. 11–18.
[4] N.K. Benamara, M. Val-Calvo, J.R. Álvarez-Sánchez, A. Díaz-Morcillo, J.M. Ferrández, E. Fernández-Jover, T.B. Stambouli, Real-time facial expression recognition using smoothed deep neural network ensemble, Integr. Comput. Aided Eng. 28 (2021) 97–111.
[5] Benitez-Quiroz, C.F., Srinivasan, R., and Martinez, A.M. (2016). EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, volume 2016-Decem, pages 5562–5570.
[6] S. Berretti, A. Del Bimbo, P. Pala, B.B. Amor, M. Daoudi, A set of selected SIFT features for 3D facial expression recognition, Proceedings - International Conference on Pattern Recognition (2010) 4125–4128.
[7] R. Biswas, V. González-Castro, E. Fidalgo, E. Alegre, Perceptual image hashing based on frequency dominant neighborhood structure applied to tor domains recognition, Neurocomputing 383 (2020) 24–38.
[8] Brock, A., Donahue, J., and Simonyan, K. (2019). Large scale GAN training for high fidelity natural image synthesis. In 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6–9, 2019. OpenReview.net.
[9] Brock, A., Lim, T., Ritchie, J., and Weston, N. (2017). Neural photo editing with introspective adversarial networks. ArXiv, abs/1609.07093.
[10] Cai, J., Meng, Z., Khan, A., Li, Z., O'Reilly, J., and Tong, Y. (2019). Identity-free facial expression recognition using conditional generative adversarial network. CoRR, abs/1903.08051.
[11] J. Cai, Z. Meng, A.S. Khan, Z. Li, J. O'Reilly, Y. Tong, Probabilistic attribute tree structured convolutional neural networks for facial expression recognition in the wild, IEEE Trans. Affect. Comput. (2022).
[12] Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. (2016). Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5–10, 2016, Barcelona, Spain, pages 2172–2180.
[13] Choi, Y., Choi, M., Kim, M., Ha, J., Kim, S., and Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018, pages 8789–8797. IEEE Computer Society.
[14] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.

[15] Z. Dai, Z. Yang, F. Yang, W.W. Cohen, R. Salakhutdinov, Good semi-supervised learning that requires a bad GAN, in: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4–9 December 2017, Long Beach, CA, USA, 2017, pp. 6510–6520.
[16] H. de Vries, F. Strub, J. Mary, H. Larochelle, O. Pietquin, A.C. Courville, Modulating early visual processing by language, in: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4–9 December 2017, Long Beach, CA, USA, 2017, pp. 6594–6604.
[17] de Vries, H., Strub, F., Mary, J., Larochelle, H., Pietquin, O., and Courville, A.C. (2017b). Modulating early visual processing by language. In NIPS.
[18] S. Deepthi, G. Archana, V. JagathyRaj, Facial Expression Recognition Using Artificial Neural Networks, IOSR Journal of Computer Engineering 8 (4) (2013) 1–6.
[19] E.L. Denton, S. Chintala, A. Szlam, R. Fergus, Deep generative image models using a laplacian pyramid of adversarial networks. In Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015(December), pp. 7–12, Montreal, Quebec, Canada, 2015, pp. 1486–1494.
[20] Ding, H., Sricharan, K., and Chellappa, R. (2018). Exprgan: Facial expression editing with controllable expression intensity. In McIlraith, S.A. and Weinberger, K.Q., editors, Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2–7, 2018, pages 6781–6788. AAAI Press.
[21] Ding, R., Guo, G., Yan, X., Chen, B., Liu, Z., and He, X. (2020). Bigan: Collaborative filtering with bidirectional generative adversarial networks. In Proceedings of the 2020 SIAM International Conference on Data Mining, SDM 2020, Cincinnati, Ohio, USA, May 7–9, 2020 [the conference was canceled because of the coronavirus pandemic, the reviewed papers are published in this volume], pages 82–90. SIAM.
[22] Dumoulin, V., Belghazi, I., Poole, B., Lamb, A., Arjovsky, M., Mastropietro, O., and Courville, A.C. (2017a). Adversarially Learned Inference. In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26, 2017, Conference Track Proceedings.
[23] Dumoulin, V., Shlens, J., and Kudlur, M. (2017b). A learned representation for artistic style. In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26, 2017, Conference Track Proceedings. OpenReview.net.
[24] P. Ekman, Basic Emotions, chapter 3, John Wiley & Sons Ltd., 1999, pp. 45–60.
[25] P. Ekman, W. Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Movement, Consulting Psychologists Press, Palo Alto, 1978.
[26] P. Ekman, W.V. Friesen, Constants across cultures in the face and emotion, J. Pers. Soc. Psychol. 17 (2) (1971) 124–129.
[27] A.P. Fard, M.H. Mahoor, Ad-corre: Adaptive correlation-based loss for facial expression recognition in the wild, IEEE Access 10 (2022) 26756–26768.
[28] Fernandez, P.D.M., Peña, F.A.G., Ren, T.I., and Cunha, A. (2019). FERAtt: Facial Expression Recognition with Attention Net. arXiv.
[29] J.M. Fernández-Dols, P. Carrera, C. Crivelli, Facial Behavior While Experiencing Sexual Excitement, J. Nonverbal Behav. 35 (1) (2011) 63–71.
[30] A. Gangwar, E. Fidalgo, E. Alegre, V. González-Castro, Pornography and child sexual abuse detection in image and video: a comparative evaluation, in: In 8th International Conference on Imaging for Crime Detection and Prevention (ICDP 2017), Institution of Engineering and Technology, 2017, pp. 37–42.
[31] A. Gangwar, V. González-Castro, E. Alegre, E. FIDALGO, Attm-cnn: Attention and metric learning based cnn for pornography, age and child sexual abuse (csa) detection in images, Neurocomputing 445 (2021) 81–104.
[32] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems 27, Curran Associates Inc., 2014, pp. 2672–2680.
[33] R. Gross, I. Matthews, J. Cohn, T. Kanade, S. Baker, Multi-PIE, in: In 2008 8th IEEE International Conference on Automatic Face & Gesture Recognition, IEEE, 2008, pp. 1–8.
[34] Haque, A. (2021). EC-GAN: Low-Sample Classification using Semi-Supervised Algorithms and GANs. ArXiv, abs/2012.15864.
[35] B. Hasani, P.S. Negi, M.H. Mahoor, Breg-next: Facial affect computing using adaptive residual networks with bounded gradient, IEEE Trans. Affect. Comput. 13 (2022) 1023–1036.
[36] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, volume 2016-December, pages 770–778. IEEE Computer Society.
[37] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, in: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4–9 December 2017, Long Beach, CA, USA, 2017, pp. 6626–6637.
[38] Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6–11 July 2015, volume 37 of JMLR Workshop and Conference Proceedings, pages 448–456. JMLR.org.
[39] Isola, P., Zhu, J., Zhou, T., and Efros, A.A. (2017). Image-to-image translation with conditional adversarial networks. In 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21–26, 2017, pages 5967–5976. IEEE Computer Society.
[40] Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2018). Progressive growing of gans for improved quality, stability, and variation. ArXiv, abs/1710.10196.
[41] Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16–20, 2019, pages 4401–4410. Computer Vision Foundation/ IEEE.
[42] Kingma, D.P. and Ba, J. (2015). Adam: A method for stochastic optimization. CoRR, abs/1412.6980.
[43] A. Krizhevsky, G. Hinton, Learning Multiple Layers of Features from Tiny Images, University of Toronto, 2009, Technical report.
[44] R.S. Lazarus, B.N. Lazarus, Passion and reason: making sense of our emotions, Oxford University Press, 1994.
[45] Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278–2324.
[46] C. Li, T. Xu, J. Zhu, B. Zhang, Triple generative adversarial nets, in: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4–9 December 2017, Long Beach, CA, USA, 2017, pp. 4088–4098.
[47] J. Li, Z. Dong, S. Lu, S.-J. Wang, W.-J. Yan, Y. Ma, Y. Liu, C. Huang, X. Fu, $CAS(ME)^3$: A Third Generation Facial Spontaneous Micro-Expression Database with Depth Information and High Ecological Validity, IEEE Transactions on Pattern Analysis and Machine Intelligence. (2022).
[48] Li, S., Deng, W., and Du, J. (2017b). Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21–26, 2017, pages 2584–2593. IEEE Computer Society.
[49] Z. Lian, Y. Li, J. Tao, J. Huang, M. Niu, Expression analysis based on face regions in real-world conditions, Int. J. Autom. Comput. 17 (2020) 96–107.
[50] Lim, J.H. and Ye, J.C. (2017). Geometric gan. ArXiv, abs/1705.02894.
[51] B. Liu, Y. Zhu, K. Song, A. Elgammal, Towards faster and stabilized gan training for high-fidelity few-shot image synthesis, International Conference on Learning (2021), Representations.
[52] Lowe, D.G. (1999). Object recognition from local scale-invariant features. In Proceedings of the Seventh IEEE International Conference on Computer Vision, volume 2, pages 1150–1157 vol 2.
[53] Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. (2010). The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010, pages 94–101.
[54] Marchesi, M. (2017). Megapixel size image creation using generative adversarial networks. ArXiv, abs/1706.00082.
[55] Meng, Z., Liu, P., Cai, J., Han, S., and Tong, Y. (2017). Identity-aware convolutional neural network for facial expression recognition. In 12th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2017, Washington, DC, USA, May 30 - June 3, 2017, pages 558–565. IEEE Computer Society.
[56] Mirza, M. and Osindero, S. (2014). Conditional generative adversarial nets. CoRR, abs/1411.1784.
[57] Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. ArXiv, abs/1802.05957.
[58] Miyato, T. and Koyama, M. (2018a). cgans with projection discriminator. In 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net.
[59] T. Miyato, M. Koyama, cgans with projection discriminator, 2018, ArXiv, abs/1802.05637.
[60] A. Mollahosseini, B. Hasani, M.H. Mahoor, AffectNet: a database for facial expression, valence, and arousal computing in the wild, IEEE Trans. Affect. Comput. 10 (1) (2019) 18–31.
[61] Y. Mroueh, S. Voinea, T. Poggio, Learning with group invariant features: A kernel perspective, in: NIPS, 2015.
[62] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, A.Y. Ng, Reading Digits in Natural Images with Unsupervised Feature Learning, In Proc, NIPS Workshop on Deep Learning and Unsupervised Feature Learning, 2011.
[63] Niu, B., Gao, Z., and Guo, B. (2021). Facial expression recognition with lbp and orb features. Computational Intelligence and Neuroscience, 2021.
[64] A. Odena, Semi-supervised learning with generative adversarial networks, 2016 ArXiv, abs/1606.01583.
[65] Odena, A., Olah, C., and Shlens, J. (2017). Conditional image synthesis with auxiliary classifier gans. In Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6–11 August 2017, volume 70 of Proceedings of Machine Learning Research, pages 2642–2651. PMLR.
[66] Orrite, C., Gañán, A., and Rogez, G. (2009). HOG-based decision tree for facial expression classification. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 5524 LNCS, pages 176–183.
[67] Pantic, M., Valstar, M., Rademaker, R., and Maat, L. (2005). Web-based database for facial expression analysis. In IEEE International Conference on Multimedia and Expo, ICME 2005, volume 2005, pages 317–321.

[68] Perarnau, G., van de Weijer, J., Raducanu, B., and Álvarez, J.M. (2016). Invertible conditional gans for image editing. CoRR, abs/1611.06355.
[69] Pumarola, A., Agudo, A., Martínez, A.M., Sanfeliu, A., and Moreno-Noguer, F. (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, volume 11214 of Lecture Notes in Computer Science, pages 835–851. Springer.
[70] Radford, A., Metz, L., and Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. In 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2–4, 2016, Conference Track Proceedings.
[71] H. Ren, A. Kheradmand, M. El-Khamy, S. Wang, D. Bai, J. Lee, Real-world super-resolution using generative adversarial networks, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 1760–1768, 2020.
[72] S. Saikia, E. Fidalgo, E. Alegre, L. Fernández-Robles, Object Detection for Crime Scene Evidence Analysis Using Deep Learning, Image Analysis and Processing - ICIAP 2017, volume 10485 LNCS, Springer Verlag, 2017, pp. 14–24.
[73] S. Saikia, E. Fidalgo, E. Alegre, L. Fernández-Robles, Query based object retrieval using neural codes, Advances in Intelligent Systems and Computing, volume 649, Springer Verlag, 2018, pp. 513–523.
[74] Salimans, T., Goodfellow, I.J., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved techniques for training gans. In Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5–10, 2016, Barcelona, Spain, pages 2226–2234.
[75] F.Z. Salmam, A. Madani, M. Kissi, Facial Expression Recognition Using Decision Trees, in: Proceedings - Computer Graphics, Imaging and Visualization: New Techniques and Trends, CGiV 2016, Institute of Electrical and Electronics Engineers Inc., 2016, pp. 125–130.
[76] A.M. Saxe, J.L. McClelland, S. Ganguli, Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. CoRR, 2014, abs/1312.6120.
[77] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7–12, 2015, pages 815–823. IEEE Computer Society.
[78] C. Shan, S. Gong, P.W. McOwan, Facial expression recognition based on Local Binary Patterns: A comprehensive study, Image Vis. Comput. 27 (6) (2009) 803–816.
[79] J. Shao, Y. Qian, Three convolutional neural network models for facial expression recognition in the wild, Neurocomputing 355 (2019) 82–92.
[80] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv, 2014.
[81] H. Siqueira, S. Magg, S. Wermter, Efficient facial feature learning with wide ensemble-based convolutional neural networks, in: AAAI, 2020.
[82] J.T. Springenberg, Unsupervised and semi-supervised learning with categorical generative adversarial networks, in: 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2–4, 2016, Conference Track Proceedings.
[83] Tran, D., Ranganath, R., and Blei, D.M. (2017). Hierarchical implicit models and likelihood-free variational inference. In NIPS.
[84] J.J. Virtusio, J.J.M. Ople, D.S. Tan, M. Tanveer, N. Kumar, K. lung Hua, Neural style palette: A multimodal and interactive style transfer from a single style image, IEEE Trans. Multimedia 23 (2021) 2245–2258.
[85] X. Wei, B. Gong, Z. Liu, W. Lu, L. Wang, Improving the improved training of wasserstein gans: A consistency term and its dual effect, in: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net.
[86] S. Wu, G. Deng, J. Li, R. Li, Z. Yu, H. Wong, Enhancing triplegan for semi-supervised conditional instance synthesis and classification, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16–20, 2019, pages 10091–10100, 2019. Computer Vision Foundation/ IEEE.
[87] H. Yan, Y. Ding, P. Li, Q. Wang, Y. Xu, W. Zuo, Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21–26, 2017, pages 945–954. IEEE Computer Society.
[88] Z. Yu, C. Zhang, Image based static facial expression recognition with multiple deep network learning, in: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, November 09–13, 2015, pages 435–442. ACM.
[89] F. Zhang, T. Zhang, Q. Mao, C. Xu, Joint Pose and Expression Modeling for Facial Expression Recognition, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2018, pages 3359–3368.
[90] H. Zhang, I.J. Goodfellow, D.N. Metaxas, A. Odena, Self-attention generative adversarial networks, in: K. Chaudhuri, R. Salakhutdinov, (Eds.), Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9–15 June 2019, Long Beach, California, USA, volume 97 of Proceedings of Machine Learning Research, pages 7354–7363. PMLR, 2019.
[91] Z. Zhang, P. Luo, C.C. Loy, X. Tang, From facial expression recognition to interpersonal relation prediction, Int. J. Comput. Vision 126 (5) (2018) 550–569.
[92] R. Zhi, M. Liu, D. Zhang, A comprehensive survey on automatic facial action unit analysis, Visual Comput. 36 (2019) 1067–1093.
[93] I. Çugu, E. Sener, E. Akbas, Microexpnet: An extremely small and fast model for expression recognition from face images, in: 2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA), 2019, pages 1–6.

Abhishek Gangwar Kumar He is a researcher in computer vision and machine learning. He received his Ph.D. in Deep Learning at the University of León. He has been a co-author in around 22 papers. He has around 15 years of industry work experience and has been involved in some big R&D projects involving theoretical as well as applied research work in machine learning and image processing etc. He has been co-inventor of 2 patents, and 5 intellectual property registries.

Dr. Víctor González-Castro Víctor González received the B.S degree in Computer Science from the University of León (Spain) in 2006 and a PhD in Computer Science from the same University in 2011. He has worked as a Postdoctoral Research Fellow at the École Nationale Supérieure des Mines de Saint-Étienne (France) between May 2013 and October 2014. After that, he moved to Edinburgh, where he worked as a Lecturer in Medical Image Analysis at the Centre for Clinical Brain Sciences of the University of Edinburgh until January 2017. He is currently an associate Professor at the Department of Electrical Engineering of the University of Leon.

Dr. Enrique Alegre Enrique Alegre received his BSc. in Industrial Engineering at the University of Cantabria and his PhD in Computer Science at the University of León. He is the director of the Research Group for Vision and Intelligent Systems of the University of León, and he has participated in 20 research projects in public and competitive calls and even a bigger number with companies, been the Principal Investigator, among others, of 3 European Projects. He has been co-inventor of 12 patents, 5 of which are licensed to companies, and 15 intellectual property registries. He has also been the co-author in more than 140 papers, 45 of which have been published in indexed journals and 33 of them in the top ones.

Dr. Eduardo Fidalgo Fernández Eduardo Fidalgo Fernández received the M. Sc. (2008) degree in Industrial Engineering and the Ph.D. degree in 2015 from the University of León. He is currently an Assistant Professor at the University of León and also the Coordinator of the Group for Vision and Intelligent Systems (GVIS) under the project (Addendum) between the University of León and INCIBE (https://www.incibe.es/en), whose objective is the research and development of solutions to problems related to cybersecurity for INCIBE, by using Artificial Intelligence. His current research interests are the application of Natural Language Processing, Computer Vision, and Machine/Deep Learning.