Triple-BigGAN: A Semi-Supervised GAN for Image synthesis and classification applied to detect facial sexual expressions

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*
[b]*Researcher at INCIBE (Spanish National Cybersecurity Institute). León, Spain*
[c]*Centre for Development of Advanced Computing (CDAC), Mumbai, 400049, India*

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 labelled 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 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 unlabelled or partially labelled data. The Triple-BigGAN framework shows promising classification performance for the SFER task (i.e., 93.59%) and other three benchmark datasets, i.e., MNIST, CIFAR-10 and SVHN. Next, we evaluated the quality of samples generated by Triple-BigGAN with

*Corresponding Author

Email addresses: abhishekg@cdac.in (Abhishek Gangwar), victor.gonzalez@unileon.es (Víctor González-Castro), ealeg@unileon.es (Enrique Alegre), efidf@unileon.es (Eduardo Fidalgo)

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 SEA-Faces-30k dataset. Further, we empirically demonstrated that synthetic erotic faces images generated by Triple-BigGAN could also help in improving the classification performance of deep supervised networks.

Keywords: Facial Expressions, Pornography, Not Safe For Work (NSFW), Obscene Image Retrieval, Deep Learning, Emotion Detection

---

1. Introduction

Automatic facial expression is an active field with multiple applications, from video surveillance to emotion-based photo capturing and tagging [70, 4, 1, 80, 23]. 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 [22]: anger, disgust, fear, happiness, sadness, and surprise. Ekman and Friesen stated that these emotions are shared across different cultures and people. However, subsequent studies argued that these six basic emotions are culture-specific and, thus, not universal [34].

Nowadays, in digital forensic, Law Enforcement Agencies (LEAs) use Multimedia-Forensic Analysis Tools (M-FAT) and Facial-Forensic Analysis Tools (F-FAT) to deal with the growing volume of data seized from cybercrimes [64, 65]. Crimes where Child Sexual Exploitation Material (CSEM) is involved are specially sensitive, due to the nature of the handled data. One of the strategies that are used to detect CSEM automatically [25] is to detect images that contain pornography and to estimate the age of people on 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.

This paper focuses on detecting sexual expressions on faces by means of automatic sexual facial expression recognition, which would make possible to verify if the image has erotic nature. This may also 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 [27, 71, 30]. However, these successes are

heavily dependent on large training sets of manually annotated data [3, 26]. As far as we know, there are not any publicly available large collections of labelled data suitable for training a deep learning model for sexual facial expression detection. Hence, first we created a manually labelled dataset containing erotic, suggestive-erotic, and non-erotic facial images. We observed that the labelling procedure in the task of sexual facial expression is much more difficult and time-consuming than normal image labelling. Motivated by aforementioned issues, in this paper we propose a Semi-Supervised Learning (SSL) framework based on Generative Adversarial Networks (GANs) [27], which can learn image representation from data from which only a small part is labelled.

In regard to learn better representations, researchers have been exploiting different methods to utilize unlabelled or partially labelled data for many years [73]. 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 labelled data. That way, such classifiers can then generalize better.

Recently, GANs have achieved an impressive success for various types of computer vision problems such as image synthesis [43], style transfer [75], image super-resolution [63] and classification [29]. In general, the 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 [16].

One interesting extension of GAN is Conditional GAN (CGAN) [48] where a condition variable can control the generated image. In an alternative approach proposed in [57], 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 [73, 56, 12, 77, 29]. Also, [78] and [73] 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 labelled data for prediction. Given that labelled 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 unlabelled.

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 labelled 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-labelled facial images. This is the first dataset of images of Sexual Expression and Activity Faces. It can be accessed through our website upon request for research purposes only[1].

- We empirically demonstrate that Triple-BigGAN provides the state-of-the-art classification accuracy on the MNIST [39], CIFAR-10 [38], SVHN [54] 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.

[1] `https://gvis.unileon.es/dataset/sea-faces/`

The description of the experiments, their results and a discussion are covered by 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) [48], 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 [48], images [33], and object/image attributes [60].

Researchers have explored different ways to convert standard GAN into CGAN [48, 73, 60, 20, 20, 57, 50]. The basic type of CGANs requires supervised information related to the condition variable(s). Springerberg [73] 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 unlabelled data [73]. Dumoulin et al. [20] and de Vries et al. [13] presented a modified class conditioning in the input to the generator by means of class conditional gains and biases in Batch Normalization layers [32]. In the work carried out by Odena et al. [57], 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 [50], the authors modified the discriminator and utilized cosine distance between its features and a set of learned class embeddings 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 [9] 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) [19] 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 [18], the authors introduced an encoder module along with 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 [40] 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, Hacque proposed the model External Classifier GAN (EC-GAN) [29], 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 [35], Karras et al. presented a new training methodology and improvements in 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 [36], extended the progressive GAN and explored multiple improve-

ments 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 [6]. 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.

Our work extends the Triple-GAN, EC-GAN, and BigGAN frameworks; however, there are significant differences with them. TripleBig-GAN follows the adversarial training methodology of Triple-GAN, however, we have redesigned the classifier, generator, discriminator and loss functions to generate high fidelity class-conditional data distributions as well as a improved classifier. The major objective of the Triple-GAN approach is to train a strong discriminator, and the EC-GAN aims mainly to train an improved classifier exploiting GAN part. Differently than both of them, our proposed approach focuses on training a strong classifier as well as a strong discriminator in an end-to-end manner. Our work differs from BigGAN since this mainly focuses on unsupervised image synthesis, whereas Triple-BigGAN aims at semi-supervised joint distribution matching. Our network can utilize labelled and unlabelled data and learns classification and synthesis, both together.

2.2. Related work in Facial Expression Recognition

Existing approaches for 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 [22] (i.e., anger, disgust, fear, happiness, sadness, and surprise), and a neutral one [55, 84, 72, 70, 23]. The other type of approaches for FER [83] involves the use of the Facial Action Coding System (FACS), which describes facial muscle movement using 44 different Action Units (AU) [21]. Each AU corresponds to a specific facial substructure, and the six basic emotions can be categorized 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) [70], Scale Invariant Feature Transform (SIFT) [44] or Histogram of Oriented Gradients (HOG) [58] followed by a classifier such as Support Vector Machine (SVM) [11], Decision Trees (DT) [67] or Artificial Neural Networks (ANN) [15]. State-of-the-art approaches for FER are based on Deep Learning, especially Convolutional Neural Network (CNN) [1, 80, 23, 72].

The hand-crafted features give good accuracy in a constraint 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 on the illumination and head pose angle.

Recently, deep neural networks have been employed to increase the robustness of FER to 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 age of subjects involved in training. The major limitation of this methodology is that it reduces the generalization of the model on the unknown identities. Despite a noticeable research in the field, modelling inter-subject differences in FER is still persisting as an open challenge.

Following this, various techniques [41, 8] 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) [47] 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 [8].

[8] 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 scheme has a challenge, which is that since FER is based on the synthetic data, its performance is influenced not only by the quality of the generated data, but also on the performance of the expression transfer between the input sample to

the synthetic sample. In [79], the authors proposed De-expression Residue Learning (DeRL) to learn subject-independent facial expression representations.

More recently, StarGAN [10] was presented to edit the facial expression and attributes. It is a multi-domain approach which learns generation of facial expressions and 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, [17] 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. [61] 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 labelled data especially, AUs.

Success in various computer vision classification problems relies heavily on the availability of the 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.

Table 1: A summary of publicly available datasets for FER

Dataset	**Num. images/ videos**	**Num. Expressions**
CK+ [45]	593 images	6 basic + neutral + contempt
FER-2013[2]	35,887 images	6 basic + neutral
MMI [59]	740 images, 2900 videos	6 basic+ neutral
Multi-PIE[28]	755,370 images	6 basic
EmotioNet [3]	1M images	23 basic or compound
AffectNet [52]	450K images	6 basic+ neutral
ExpW [82]	91,793 images	6 basic+ neutral

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 have 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 [2] 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 analysed some common behaviour 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. [24] 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 [2].

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 not-erotic facial images from people without any sexual activity.

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

SEA-Faces-30k contains 30,817 images collected from the Internet in which the faces show any kind of sexual expression. The dataset has three categories based on the intensity of adult content: erotic, suggestive-erotic, and non-erotic. Face information changes depending on the age, sex, face shape, skin or 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 artefacts on the face such as glasses, hats, beards, or jewellery. 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. The main features of SEA-Faces-30k are summarized in Table 2. Figures 1, 2 and 3 de-

pict some examples of the erotic, suggestive-erotic, and non-erotic categories, respectively.

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

Category	**Num. images**
Erotic	10,399
Suggestive-Erotic	10,160
Non-Erotic	10,258
TOTAL IMAGES	30,817

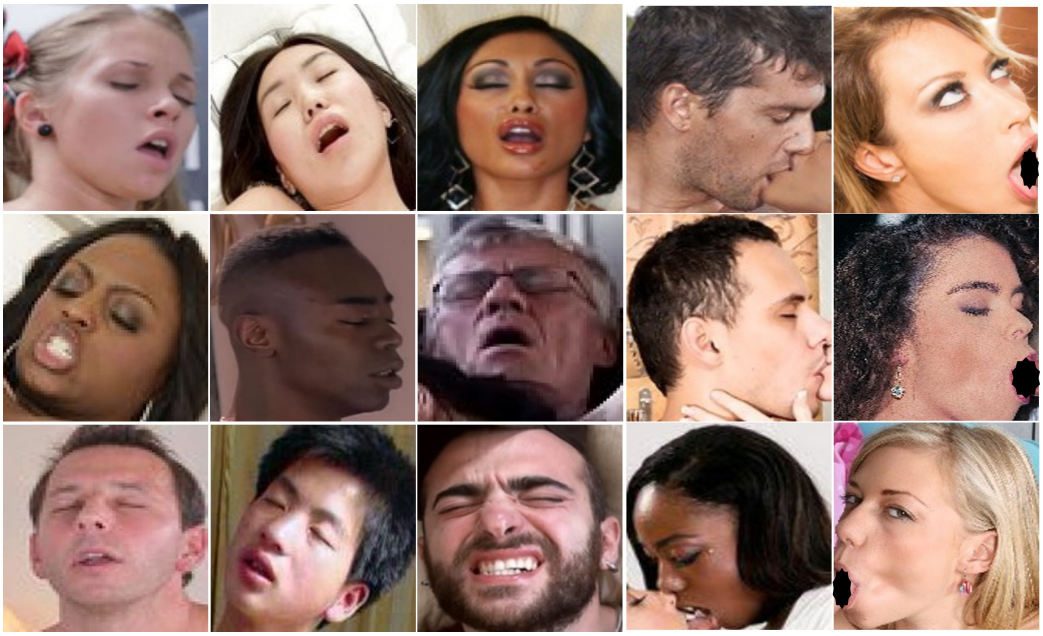


Figure 1: Examples of images from the class "erotic" of SEA-Faces-30k.

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

During the *data crawling*, we gathered around 50k pornographic images from two popular pornographic websites. Due to the varied content of the images, e.g. from a single clothed model posing to nude group sexual ac-

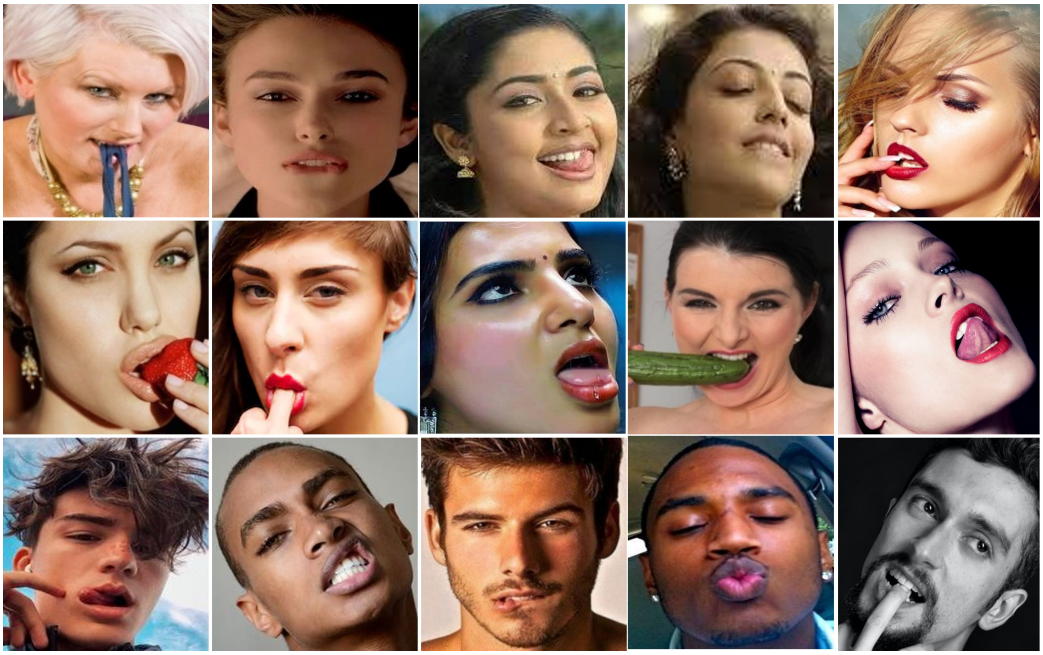


Figure 2: Examples of images from the class "suggestive-erotic" in SEA-Faces-30k.

tivity, we applied the Yahoo open NSFW pornography detector[3], which assigns a score to each image depending on the pornographic content (i.e., the higher the score, the higher the pornographic content). Then, the crawled images are classified 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 possible classification errors were corrected manually in further stages.

We also noticed that in the crawled data the many images were from similar people and captured in controlled environment. Therefore, in order to have more variability and general web images in the data, we used Google search engine to retrieve 10k additional images in each of the three categories. For the *erotic* class, we used the following 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", "horny". Finally,

[3] `https://github.com/yahoo/open_nsfw`

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

to enhance the *non-erotic* facial data, 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".

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. To remove such duplicated images, in the stage of *removal of duplicated images*, we used a perceptual hashing [5] method, i.e. pHash[4] 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 means completely similar and towards 64 represents more dissimilarity. We empirically found that a threshold of four is good enough to detect duplicate or near-duplicate images in our data.

[4] `https://www.phash.org`

Next, in the step of *Face detection and alignment*, we used the RetinaFace detector[5], for detecting the faces on the images, as well as for getting the following landmark points: left and right eye centres, left and right mouth corners and nose tip. These landmark points are used for face alignment. We only selected the face images with resolution higher than 50×50 pixels.

During the *Manual filtering* step, we removed all the images in which either a human face was not present (i.e. due to an error of the face detector), or it was not recognizable by a human due to very bad quality, occlusions, or large pose.

Finally, we started the *manual labelling and categorization* stage, where the facial images belonging to the respective category were selected through visual inspection. It should be noted that the facial images in which expressions were not visible properly to decide them as erotic or suggestive-erotic were kept in the non-erotic category.

4. Proposed approach

The GAN framework introduced in this paper is an extension of previously proposed GANs: Conditional GAN (CGAN) [48], Semi-Supervised GAN (SSL-GAN) [73, 56, 12, 77], Triple-GAN [40], EC-GAN [29], and BigGAN [6]. 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:

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

[5] `https://github.com/deepinsight/insightface/tree/master/RetinaFace`

where $\mathbb{E}_x$ represents the expected value over all the data samples. The conditional generative adversarial network [48] 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:

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

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 [73]. As a result, unlabelled data can be leveraged to learn a good representation. In [62], 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 labelled data. In most previous works, to extend standard GANs to semi-supervised GAN (i.e. to utilize labelled and unlabelled data), the discriminator output is modified to have K outputs corresponding to real classes [73]. In some works, an additional $(K+1)^{th}$ class corresponding to the fake data generated by the generator is added [66, 19], 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 [53]. 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 [40], 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 unlabelled data and also the supervision from the classification loss on the labelled 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 [6], 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 synthetic images. Through various experiments, the paper demonstrates that the strategies of increasing the batch size and use 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 [81], and it also adopted hinge loss [74, 42] as adversarial loss function, which is similar to SAGAN. Furthermore, the authors adopted the Truncation Trick, originally proposed in [46], and Off-Diagonal Orthogonal Regularization, which is a variant of the Orthogonal Regularization proposed in [7]. To utilize class information in the generator, the BigGAN exploited the class-conditional batch normalization [14], and to utilize it in the discriminator they adopted projection discriminator [51]. The BigGAN optimization follows the SAGAN guidelines [81] and employs the Spectral Normalization [49], however, different from SAGAN, BigGAN took two steps of discriminator per generator step. To initialize the latent vector z, the orthogonal initialization approach [68] 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

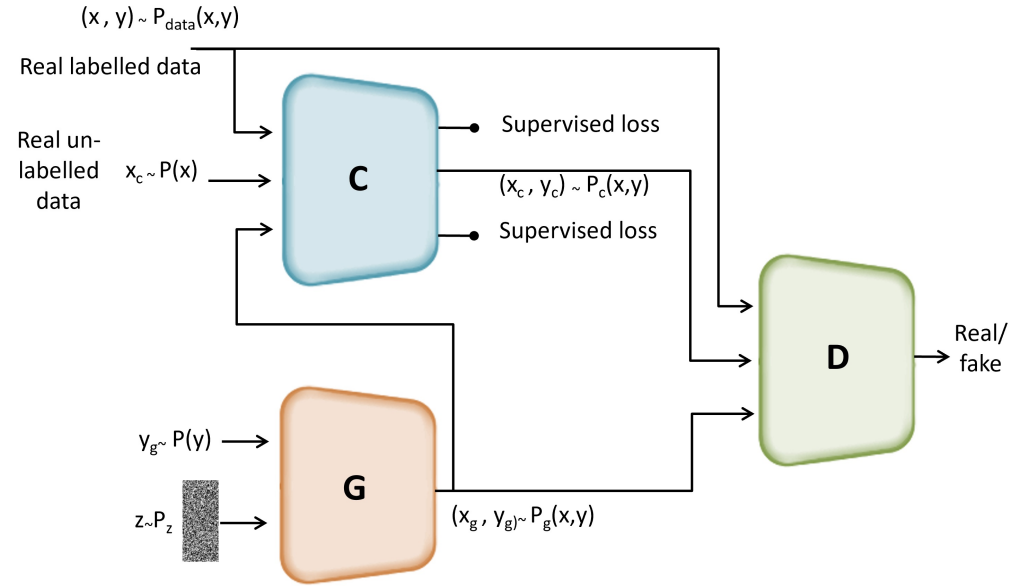
We will first formulate the semi-supervised learning setting adopted in this paper. We denote the images in the dataset as $X = \{x_1, ..., x_N\}$, $x_i \in \mathbb{R}^{r \times c}$. Let us assume our dataset contains N images, out of which N_L images contain ground-truth class labels $y_i \in \{1, 2, ..., K\}$, 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 labelled, i.e. $N_L > 0$ and $N_L << N_{UL}$.

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 easily available large amount of unlabelled data (e.g. faces extracted from pornographic and non-pornographic images unlabelled for facial expressions) to improve the performance of the model for the target problem when the labelled 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 Figure 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 an 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 the 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., labelled as well as unlabelled data (P_{data}) to learn $G(z, y)$, which can generate samples similar to $P(x|y)$, and for this

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



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 labelled and unlabelled data. The labelled data will be used to provide supervision for the classifier, whereas the unlabelled 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 unlabelled 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 which can output labels for the data accurately enough to consider them equivalent to the ground truth labels, i.e. $P_c(x, y) \sim P_{data}$, and to

learn a generator which can generate synthetic call-conditional samples similar to the true data distribution, i.e. $P_g(x, y) \sim P_{data}$. The whole network attain 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 unlabelled 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 unlabelled 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 generator and classifier. In our experiments, we assigned the same weights to the classifier and generator.

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 dataset. Finally, we evaluated the classification performance on three general-purpose benchmark datasets, i.e., MNIST [39], CIFAR-10 [38] and SVHN [54]. Furthermore, we also performed an additional experiment to evaluate the use-

fulness of synthetic images generated by Triple-BigGAN to improve the accuracy of deep CNN networks by augmenting the training data.

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 Street View House Numbers (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 case of the MNIST, SVHN and CIFAR-10 datasets, we adopted the same settings that have been adopted by many previous works [77, 40, 76, 19, 73, 12]. Specifically, we performed experiments for the cases in which there are 100, 1000, and 4000 randomly selected labelled test instances, respectively. On each of these cases, the 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 [77], Triple-GAN [40], CT-GAN [76], ALI [19], CatGAN [73] and GoodBadGAN [12].

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 5000 labelled images from the training set to understand its capabilities with 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 [69], and two famous deep CNNs, i.e., VGG-16 [71] and ResNet-50 [30]. 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[6], 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[7]. 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) Layers 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) [66] and the Frechet Inception Distance (FID) [31]. 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 (12GB) and two Tesla K80 GPUs (24GB).

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 3 and 4, respectively.

[6] `http://www.cbsr.ia.ac.cn/english/CASIA-WebFace-Database.html`

[7] `https://github.com/d-acharya/CovPoolFER`

Tables 5 and 6 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 represents 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 down sampling) does not contain the skip connection layer. In the classifier, the network adopted is ResNet-50 [30] 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.

Table 3: Architecture for Triple-BigGAN's Generator. Note that "ch" is the channel width multiplier (i.e. $ch = 128$ in Triple-BigGAN). "BN" stands for the batch normalization and "SN" denotes the 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, 3×3 Conv $ch \rightarrow 3$	-	$256 \times 256 \times 3$
Tanh	-	$256 \times 256 \times 3$

The training of our Triple-BigGAN network follows the guidelines for the SAGAN [81] and BigGAN [6] networks. The weights in Triple-BigGAN's generator, discriminator, and classifier networks are initialized through orthogonal initialization [68]. We used the Adam [37] optimizer (momentum parameters $\beta_1 = 0$, $\beta_2 = 0.999$) with the learning rates $5 \cdot 10^{-5}$ for the generator, $2 \cdot 10^{-4}$ for the discriminator, and $2 \cdot 10^{-4}$ for the classifier. We also utilized spectral normalization [49] and Orthogonal Regularization [7] in the generator and discriminator (but not in the classifier) for training stability. In the discriminator, the spectral normalization is used in all weight layers. The random noise as input to the generator is drawn from the normal distribution $\mathcal{N}(0, I)$. During training, we performed two discriminator steps for each generator and classifier step and training is done for 500k steps with a

Table 4: 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 $3 \to ch$	$128 \times 128 \times 1 \cdot ch$
ResBlock down $ch \to ch$	$64 \times 64 \times 1 \cdot ch$
Non-Local Block	$64 \times 64 \times 1 \cdot ch$
ResBlock down $ch \to 2 \cdot ch$	$32 \times 32 \times 2 \cdot ch$
ResBlock down $2 \cdot ch \to 4 \cdot ch$	$16 \times 16 \times 4 \cdot ch$
ResBlock down $4 \cdot ch \to 8 \cdot ch$	$8 \times 8 \times 8 \cdot ch$
ResBlock down $8 \cdot ch \to 16 \cdot ch$	$4 \times 4 \times 16 \cdot ch$
ResBlock $16 \cdot ch \to 16 \cdot ch$	$4 \times 4 \times 16 \cdot ch$
ReLU, Global sum pooling	$1 \times 1 \times 16 \cdot ch$
Embed(y)$\cdot h + (dense \to 1)$	1

Table 5: Architecture of the Residual Block in Triple-BigGAN's Generator (i.e., ResBlock up in Table 3).

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	-	$2 \cdot H \times 2 \cdot W \times C_{out}$

batch size of 256. Due to memory constraint, 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 translation by 20% and horizontal flip. All the images are randomly rotated in the range of $\pm 30°$. Therefore, the number of images obtained are 4 (scales) $\times$ 4 (translations) $\times$ 2 (flip) = 32 times the original images. Thereafter, the images were resized

Table 6: Architecture of Residual Block in Triple-BigGAN's Discriminator (i.e., ResBlock down in Table 4).

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	-	$H/2 \times W/2 \times C_{out}$

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 the Figures 5 and 6, respectively.

It can be noticed that our model can synthesise images with high-quality and large variety in 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 7, i.e., Triple-GAN [40], SAGAN [81], BigGAN [6], and StyleGAN [36]. 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 Figure 7. We observed that the significant mistakes in the generated samples were local, i.e., mainly artefacts, 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.

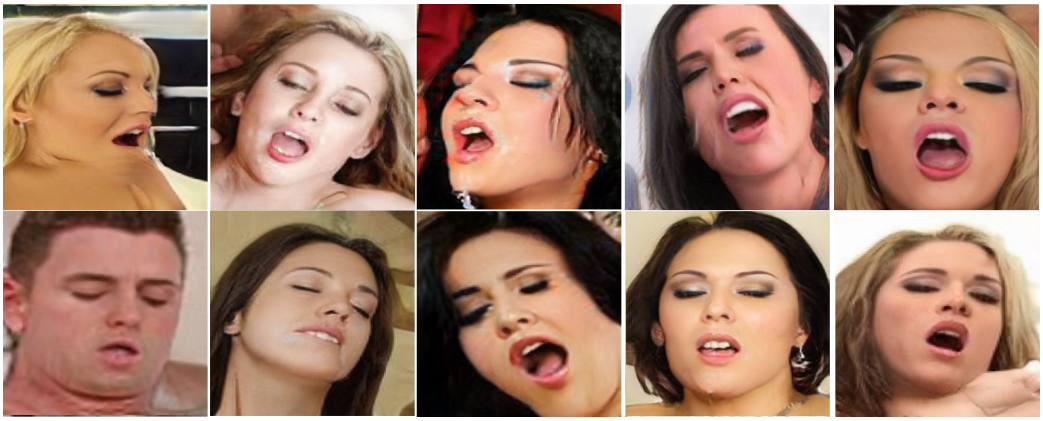


Figure 5: Samples of images from the class "erotic" generated By Triple-BigGAN. Their resolution was 256×256 pixels

Table 7: 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**

5.4. Triple-BigGAN for Classification: Sexual Facial Expression Recognition

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%, 20% and 10% of the images of the dataset, respectively.

For comparative analysis, we utilized three handcrafted descriptor-based approaches, i.e., LBP [70], HOG [58], SIFT [44], two Deep CNN networks, i.e., VGG-16 [71] and ResNet-50 [30], a deep learning-based FER approach, i.e., CovPoolFER [1], and four inference based SSL GANs, i.e., ALI [19], CatGAN [73], Triple-GAN [40] and GoodBadGAN [12]. Following the pre-

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

viously adopted protocols in [40, 19, 66], we performed the evaluation under two settings: a) with 3000 labelled images (i.e., 1000 from each category) and, b) when using all the labelled images in SEA-Faces-30k dataset.

Table 8 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 the evaluation settings, i.e., 15.77% when only 3000 labelled examples were used, and 6.42% when all the labelled 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 unlabelled data along with labelled 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 unlabelled data in addition to the labelled 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

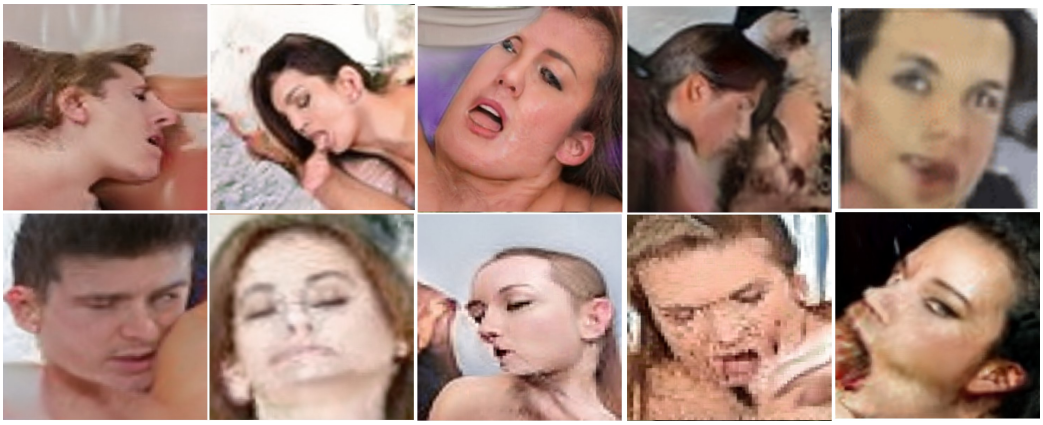


Figure 7: Samples of wrong images generated by Triple-BigGAN. Their resolution was 256×256 pixels

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%.

Table 8: Comparative analysis of Triple-BigGAN for Sexual Facial Expression Recognition. ULD refers to whether the method uses unlabelled 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**

5.5. Triple-BigGAN generated labelled 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 9. 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.

Table 9: 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

5.6. Triple-BigGAN for 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 evaluation of classification. Following the evaluation settings widely adopted [40, 77, 76, 19, 73, 12] 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 results. The error rates of this classification experiment are presented in Table 10. From the Table 10, 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.

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,

Table 10: 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

we investigated if the facial information can be utilised 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 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 approached. Our approach not only obtained a remarkable accuracy of 93.59% for 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 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 the three datasets.

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.

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

7. References

References

[1] Acharya, D., Huang, Z., Paudel, D. P., and Van Gool, L. (2018). 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] Basson, R. (2015). Human sexual response. In Handbook of Clinical Neurology, volume 130, pages 11–18. Little, Brown.

[3] 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.

[4] Berretti, S., Del Bimbo, A., Pala, P., Amor, B. B., and Daoudi, M. (2010). A set of selected SIFT features for 3D facial expression recognition. In Proceedings - International Conference on Pattern Recognition, pages 4125–4128.

[5] Biswas, R., González-Castro, V., Fidalgo, E., and Alegre, E. (2020). Perceptual image hashing based on frequency dominant neighborhood structure applied to tor domains recognition. Neurocomputing, 383:24–38.

[6] 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.

[7] Brock, A., Lim, T., Ritchie, J., and Weston, N. (2017). Neural photo editing with introspective adversarial networks. ArXiv, abs/1609.07093.

[8] 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.

[9] 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.

[10] 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.

[11] Cortes, C. and Vapnik, V. (1995). Support-Vector Networks. Machine Learning, 20(3):273–297.

[12] Dai, Z., Yang, Z., Yang, F., Cohen, W. W., and Salakhutdinov, R. (2017). 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, pages 6510–6520.

[13] de Vries, H., Strub, F., Mary, J., Larochelle, H., Pietquin, O., and Courville, A. C. (2017a). 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, pages 6594–6604.

[14] de Vries, H., Strub, F., Mary, J., Larochelle, H., Pietquin, O., and Courville, A. C. (2017b). Modulating early visual processing by language. In NIPS.

[15] Deepthi, S., Archana, G., and JagathyRaj, V. (2013). Facial Expression Recognition Using Artificial Neural Networks. IOSR Journal of Computer Engineering, 8(4):1–6.

[16] Denton, E. L., Chintala, S., Szlam, A., and Fergus, R. (2015). 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 7-12, 2015, Montreal, Quebec, Canada, pages 1486–1494.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] Ekman, P. and Friesen, W. (1978). Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, Palo Alto.

[22] Ekman, P. and Friesen, W. V. (1971). Constants across cultures in the face and emotion. Journal of Personality and Social Psychology, 17(2):124–129.

[23] Fernandez, P. D. M., Peña, F. A. G., Ren, T. I., and Cunha, A. (2019). FERAtt: Facial Expression Recognition with Attention Net. arXiv.

[24] Fernández-Dols, J. M., Carrera, P., and Crivelli, C. (2011). Facial Behavior While Experiencing Sexual Excitement. Journal of Nonverbal Behavior, 35(1):63–71.

[25] Gangwar, A., Fidalgo, E., Alegre, E., and González-Castro, V. (2017). Pornography and child sexual abuse detection in image and video: a comparative evaluation. In 8th International Conference on Imaging for Crime Detection and Prevention (ICDP 2017), pages 37–42. Institution of Engineering and Technology.

[26] Gangwar, A., González-Castro, V., Alegre, E., and FIDALGO, E. (2021). Attm-cnn: Attention and metric learning based cnn for pornography, age and child sexual abuse (csa) detection in images. Neurocomputing, 445:81–104.

[27] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In Advances in Neural Information Processing Systems 27, pages 2672–2680. Curran Associates, Inc.

[28] Gross, R., Matthews, I., Cohn, J., Kanade, T., and Baker, S. (2008). Multi-PIE. In 2008 8th IEEE International Conference on Automatic Face & Gesture Recognition, pages 1–8. IEEE.

[29] Haque, A. (2021). EC-GAN: Low-Sample Classification using Semi-Supervised Algorithms and GANs. ArXiv, abs/2012.15864.

[30] 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.

[31] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). 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, pages 6626–6637.

[32] 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.

[33] 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.

[34] Jack, R. E., Garrod, O. G. B., Yu, H., Caldara, R., and Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. Proceedings of the National Academy of Sciences of the United States of America, 109(19):7241–4.

[35] Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2018). Progressive growing of gans for improved quality, stability, and variation. ArXiv, abs/1710.10196.

[36] 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.

[37] Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. CoRR, abs/1412.6980.

[38] Krizhevsky, A. and Hinton, G. (2009). Learning Multiple Layers of Features from Tiny Images. Technical report, University of Toronto.

[39] 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.

[40] Li, C., Xu, T., Zhu, J., and Zhang, B. (2017a). 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, pages 4088–4098.

[41] 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.

[42] Lim, J. H. and Ye, J. C. (2017). Geometric gan. ArXiv, abs/1705.02894.

[43] Liu, B., Zhu, Y., Song, K., and Elgammal, A. (2021). Towards faster and stabilized {gan} training for high-fidelity few-shot image synthesis. In International Conference on Learning Representations.

[44] 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.

[45] 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.

[46] Marchesi, M. (2017). Megapixel size image creation using generative adversarial networks. ArXiv, abs/1706.00082.

[47] 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.

[48] Mirza, M. and Osindero, S. (2014). Conditional generative adversarial nets. CoRR, abs/1411.1784.

[49] Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. ArXiv, abs/1802.05957.

[50] 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.

[51] Miyato, T. and Koyama, M. (2018b). cgans with projection discriminator. ArXiv, abs/1802.05637.

[52] Mollahosseini, A., Hasani, B., and Mahoor, M. H. (2019). AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. IEEE Transactions on Affective Computing, 10(1):18–31.

[53] Mroueh, Y., Voinea, S., and Poggio, T. (2015). Learning with group invariant features: A kernel perspective. In NIPS.

[54] Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. (2011). Reading Digits in Natural Images with Unsupervised Feature Learning. In Proc. NIPS Workshop on Deep Learning and Unsupervised Feature Learning.

[55] Niu, B., Gao, Z., and Guo, B. (2021). Facial expression recognition with lbp and orb features. Computational Intelligence and Neuroscience, 2021.

[56] Odena, A. (2016). Semi-supervised learning with generative adversarial networks. ArXiv, abs/1606.01583.

[57] 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.

[58] 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.

[59] 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.

[60] Perarnau, G., van de Weijer, J., Raducanu, B., and Álvarez, J. M. (2016). Invertible conditional gans for image editing. CoRR, abs/1611.06355.

[61] 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.

[62] 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.

[63] Ren, H., Kheradmand, A., El-Khamy, M., Wang, S., Bai, D., and Lee, J. (2020). Real-world super-resolution using generative adversarial networks. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 1760–1768.

[64] Saikia, S., Fidalgo, E., Alegre, E., and Fernández-Robles, L. (2017). Object Detection for Crime Scene Evidence Analysis Using Deep Learning. In Image Analysis and Processing - ICIAP 2017, volume 10485 LNCS, pages 14–24. Springer Verlag.

[65] Saikia, S., Fidalgo, E., Alegre, E., and Fernández-Robles, L. (2018). Query based object retrieval using neural codes. In Advances in Intelligent Systems and Computing, volume 649, pages 513–523. Springer Verlag.

[66] 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.

[67] Salmam, F. Z., Madani, A., and Kissi, M. (2016). Facial Expression Recognition Using Decision Trees. In Proceedings - Computer Graphics, Imaging and Visualization: New Techniques and Trends, CGiV 2016, pages 125–130. Institute of Electrical and Electronics Engineers Inc.

[68] Saxe, A. M., McClelland, J. L., and Ganguli, S. (2014). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. CoRR, abs/1312.6120.

[69] Schroff, F., Kalenichenko, D., and Philbin, J. (2015). 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.

[70] Shan, C., Gong, S., and McOwan, P. W. (2009). Facial expression recognition based on Local Binary Patterns: A comprehensive study. Image and Vision Computing, 27(6):803–816.

[71] Simonyan, K. and Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv.

[72] Siqueira, H., Magg, S., and Wermter, S. (2020). Efficient facial feature learning with wide ensemble-based convolutional neural networks. In AAAI.

[73] Springenberg, J. T. (2016). 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.

[74] Tran, D., Ranganath, R., and Blei, D. M. (2017). Hierarchical implicit models and likelihood-free variational inference. In NIPS.

[75] Virtusio, J. J., Ople, J. J. M., Tan, D. S., Tanveer, M., Kumar, N., and lung Hua, K. (2021). Neural style palette: A multimodal and interactive style transfer from a single style image. IEEE Transactions on Multimedia, 23:2245–2258.

[76] Wei, X., Gong, B., Liu, Z., Lu, W., and Wang, L. (2018). 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.

[77] Wu, S., Deng, G., Li, J., Li, R., Yu, Z., and Wong, H. (2019). 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. Computer Vision Foundation / IEEE.

[78] Yan, H., Ding, Y., Li, P., Wang, Q., Xu, Y., and Zuo, W. (2017). 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.

[79] Yu, Z. and Zhang, C. (2015). 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.

[80] Zhang, F., Zhang, T., Mao, Q., and Xu, C. (2018a). Joint Pose and Expression Modeling for Facial Expression Recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 3359–3368.

[81] Zhang, H., Goodfellow, I. J., Metaxas, D. N., and Odena, A. (2019). Self-attention generative adversarial networks. In Chaudhuri, K. and Salakhutdinov, R., editors, 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.

[82] Zhang, Z., Luo, P., Loy, C. C., and Tang, X. (2018b). From Facial Expression Recognition to Interpersonal Relation Prediction. International Journal of Computer Vision, 126(5):550–569.

[83] Zhi, R., Liu, M., and Zhang, D. (2019). A comprehensive survey on automatic facial action unit analysis. The Visual Computer, 36:1067–1093.

[84] Çugu, I., Sener, E., and Akbas, E. (2019). Microexpnet: An extremely small and fast model for expression recognition from face images. 2019

Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA), pages 1–6.