International Journal of Multimedia Information Retrieval (2023) 12:26
https://doi.org/10.1007/s13735-023-00285-6

REGULAR PAPER

Neural style transfer generative adversarial network (NST-GAN) for facial expression recognition

Faten Khemakhem[1,2] · Hela Ltifi[1,3]

Received: 27 December 2022 / Revised: 12 May 2023 / Accepted: 10 July 2023 / Published online: 23 August 2023

Abstract
With the increasing number of intelligent human–computer systems, more and more research is focusing on human emotion recognition. Facial expressions are an effective modality in emotional recognition, enhancing automatic emotional analysis. Although significant studies have investigated automatic facial expression recognition in the past decades, previous works were mainly produced for controlled environments. Unlike recent pure CNN-based works, we argue that it is practical and feasible to recognize an expression from a facial image. However, the extracted features may capture more identity-related information and are not purely associated with the specific task of expression recognition. To reduce the personal influence of identity-related features by removing identity information from facial images, we propose a neural style transfer generative adversarial network (NST-GAN) in this paper. The objective is to determine the expression information from the input image by removing identity information and transferring it to a synthetic identity. We employ experimental strategies to evaluate the proposed method on three public facial expression databases (CK+, FER-2013, and JAFFE). Extensive experiments prove that our NST-GAN outperforms other methods, setting a new state of the art.

Keywords Facial expression recognition · Identity-free expression · Neural style transfer generative adversarial network

1 Introduction

Facial expression plays a vital role in nonverbal communication and human interactions. FER has crucial importance in developing interactive computing systems. An expression presents a remarkable signal that human beings use, purposefully or unintentionally, to transfer a message, such as an emotional state or a health condition. A study by Ekman and Friesen [1] showed that the human manner of expressing emotion is universal and assumed to be a physiological phenomenon, not depending on a particular culture. This discovery gives the current Computer Vision searches to well focus on the expression as a signal to recognize the facial message (emotion). This leads rise to various systems in different areas, going from Human–robot interaction to Data Analysis [2].

As an emerging research topic to develop an advanced human–robot interaction (HRI) system, facial expression is supposed to present physical and social contact between human beings and robots and emotional interaction. Furthermore, various machine learning algorithms have been proposed specifically to automate FER [2–4]. Effective solutions to solve FER problems have been thoroughly researched. A FER approach aims to classify a single image face as one of the six main emotional expressions [1], viz. anger, disgust, fear, happiness, sadness, and surprise, and one neutral. Moreover, recognizing facial expressions from videos is a relevant and current issue in this topic of research. On the one hand, more challenges arise in video data than image data, e.g., the variable and quick dynamics among, the beginning session, its apex, and its disappearance. On the other hand, the quantity of information produced by the set of frames or by the associated speech in some cases allows various approaches, e.g., [5–7]. These works grant

✉ Faten Khemakhem
f.khemakhem@mac.gov.tn

✉ Hela Ltifi
hela.ltifi@ieee.org

[1] REsearch Groups in Intelligent Machines, National School of Engineers, University of Sfax, Sfax, Tunisia

[2] Department of Computer Sciences, Faculty of Economics and Management, University of Sfax, Sfax, Tunisia

[3] Department of Computer Sciences, Faculty of Sciences and Techniques of Sidi Bouzid, University of Kairouan, Kairouan, Tunisia

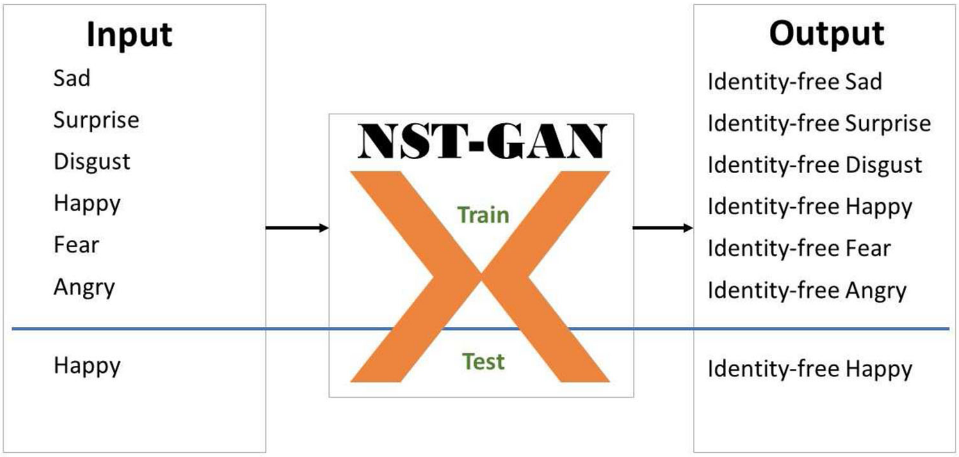

Fig. 1 The entire framework of the NST-GAN generates identity-free expression from an input image

the extraction features and classification in a robust manner and multi-modalities setting.

However, the extracted features can capture more identity-related information and thus are not strictly related to the intended task, i.e., facial expression recognition. The challenge is to reduce the personal effects of identity-accorded features by taking out identity facts from facial images. In addition, we suppose *fun(I)* to be the obtained knowledge of an input facial image, named *I*, learned by CNN. *fun(I)* is generally a nonlinear function that presents the sum of two attributes, i.e., *fun(I) = g(fid (I); fexp (I))*, whose *fid (I)* presents the identity features related by race, age, gender, and identity face; and *fexp (I)* presents expression-related information only. The purpose of the proposed system is to obtain the best exploitation of *fexp (I)* while reducing the influence of *fid (I)*.

To meet this goal, we present a novel FER approach based on neural style transfer generative adversarial network, as shown in Fig. 1, that learns emotion without identity information. We called our model NST-GAN.

The objective is to determine the expression information from the input image by removing identity information and transferring it to a synthetic "average" identity. Average expressive images are generated from all input data in the trained dataset. Hence, these images will be considered for expression classification. In summary, the following are the major contributions:

- Developing an NST-GAN pattern to modify the facial image to an average identity for expression recognition, and the recognition with adversarial changes on facial images can achieve more performance than those on natural images. That is, the reduction in perturbation is required to obtain successful recognition.
- We experimentally prove that our approach attains a considerable performance compared to the state-of-the-art FER systems.

The rest of the paper is organized as follows: Sect. 2 delivers the related works. The next section presents the proposed method. The experimental settings and results are given in Sect. 4. The quantitative and qualitative evaluation of the proposed method is discussed. In the end, we have concluded our observation of future works in Sect. 5.

2 Related works

The first works in this field employ traditional practices. Accordingly, notable studies often use handcrafted feature extraction and classification. Luo et al. [8] present a hybrid technique based on Principal Component Analysis and Local Binary Pattern (LBP). The first component is employed to extract the global features of an image. The second component is applied to extract local features. Thus, the Support Vector Machine (SVM) is used for expression classification. Chen et al. [9] applied Histograms of Oriented Gradients (HOG) to transform facial components as features. To perform the facial expression classification, SVM was applied. They evaluated the proposed method on small datasets (JAFFE and extended Cohn-Kanade (CK+)).

Although classical machine learning approaches, based on handcrafted feature extraction, have proved to predict facial expressions in controlled conditions. Recent works have revealed that these algorithms are inflexible to predict images taken in an unsupervised environment [10]. Nevertheless, these algorithms are far from generalization capacity. The main weakness of these approaches derives essentially from the fact that these techniques are only able to recognize limited or exaggerated facial expressions corresponding to the existing training set. Major factors that play a harder problem are the following: face orientation, head pose variation, irregular nature of the human face, and illumination conditions. The challenge is to extract robust features from the facial image and yet preserve the expressive information.

Recent research investigated extracting and learning features automatically from data [11]. Convolution Neural Network (CNN) is the most prevalent pattern used for image prediction. Furthermore, new experimental approaches have been developed for their use in the current intelligent systems [12]. Nevertheless, the results achieved are impressive and can be granted to the effective use of GPUs, dropout, ReLUs, and data augmentation techniques [13]. As a well-known Deep Learning (DL) pattern, CNN has thus been inspired by human beings' innate visual perception mechanism [14].

It is underlined that searches in this field are hindered by the need for a very large training set, typically required in current DL approaches. Indeed, existing FER datasets generally have a few numbers of subjects, slight variations between sets, or limited sample images per expression, hampering the training procedure. For example, one of the largest FER datasets, FER2013, has 35,887 images of seven subjects. However, only 547 of them present a disgusting portrait. Collecting and annotating a new database is often a hard, time-consuming, and expensive task. Finding alternative techniques is becoming a challenge to enhancing FER systems' performance.

Considering the disadvantages of handcrafted features, deep learning took up the challenge in this field. Xie et al. [15] proposed a new FER framework based on different feature sparseness-based regularization strategies. The feature sparseness of the hidden units is integrated into a simple convolution neural network to enhance the discrimination generalization ability. A deep metric learning [16] framework is used to optimize the regularization, which is integrated into the loss function. This technique requires a large number of parameters for optimization. The sparseness is embedded directly in the FC layers to detect common features among different people.

Jain et al. [17] propose a Hybrid Convolution-Recurrent Neural Network. Their model employed Convolution layers followed by a Recurrent Neural Network (RNN). CNN primarily is employed for feature extraction. The temporal dependencies are detected from facial images using the RNN. The relation within images is considered during the classification. During classification, the relation within images is considered. The hybrid model achieved greater performance as compared to a single CNN. Using Relu as an activation function, the major disadvantage of RNN is that the training model is unstable.

Kumar et al. [18] very recently proposed a new deep CNN, which contains convolution layers and deep residual blocks. Their model achieved a high accuracy compared with other works [8]. However, their proposed method was evaluated on only two small databases.

There are several techniques for automatic expression recognition systems in the literature as shown in Table 1. These techniques typically involve face detection via camera sensors, feature extraction, and emotional state classification, where the second step is the most crucial. The classification accuracy is mainly dependent on the pertinent extracted features.

In this work, we focus on classifying a single face from a static image, rather than a video record. There are many challenges arising in video data that make image classification less complex. Nevertheless, current works show that facial expression prediction on an image is an active and advanced research area in this field and may have a significant impact on emotion recognition from videos too.

3 Proposed method

In this work, we focus on NST-GAN, which may achieve high purity distinction between expression-related and identity-related variables. We introduce our proposed strategy based on the GAN mechanism in this section. The following sections provide a brief review of the proposed approach.

3.1 A brief review of generative adversarial networks

GANs are recently achieving majestic results in image generation [19] and document enhancement [20]. This section investigates the several integrations of this technique in related problems to facial expression image processing and enhancement. Recently, some researchers have exploited the GAN in FER problems.

Zhang et al. [21] presented a novel end-to-end deep learning approach by exploiting pose-invariant facial expression recognition. Their model can generate different expressions from face images under arbitrary poses to enrich the training set. De-Expression Residue Learning (DeRL) was introduced by Yang et al. [22] to recognize facial expressions by extracting information from the expressive component. This model automatically generates an appropriate neutral face for an input image.

GAN presents an example of an algorithm-level competitive parallel model. Traditional GANs [23] employ deep neural architectures to produce realistic images in order to perform an intelligent system involving two deep neural networks (DNNs): a generator "*G*", and a discriminator "*D*". The two DNNs are met in a zero-sum game. In other words, if one wins, the other loses. The generator generates fake input data to mislead the discriminator. While the discriminator trains to make out between fake and real samples. The discriminator and generator networks are learned simultaneously as adversaries. Two networks are evolving in parallel following their optimization process.

Table 1 Comparison of classification approaches in related works

Author(s)/references	FER approaches	Expressions	Advantages	Drawbacks
Luo et al. [8]	Hybrid technique based on Principal Component Analysis and LBP	Angry, disgust, fear, happy, sad, and surprise	Reduce computational cost and memory cost with the gain super result	Evaluated only on one small dataset in a controlled environment
Chen et al. [9]	Histograms of Oriented Gradients (HOG) and SVM	Angry, disgust, fear, happy, sad, surprise, and neutral	Characterizing the shapes of important components constitutes facial expressions	Evaluated only on two small datasets. In a controlled environment
Xie et al. [15]	A feature sparseness-based regularization that learns deep features is used. The regularization is embedded into the loss function with a simple network	Angry, disgust, fear, happy, sad, surprise, and neutral	A simple network with the proposed sparseness outperforms the one with the L2-norm regularization	Requires a large number of parameters for optimization
Jain et al. [17]	Hybrid CNN-RNN model	Angry, disgust, fear, happy, sad, surprise, and neutral	Reduces the false detection of the model	Using Relu as an activation function in RNN, the training model is unstable
Kumar et al. [18]	Deep CNN which contains convolution layers and deep residual blocks	Angry, disgust, fear, happy, sad, and surprise	A single Deep CNN	Evaluated only on two small datasets

D and *G* play the following two-player mini-max game with the following value function *V(G,D)*:

$$\min_G \max_D V(D, G) = E[\log(D(x))] + E[\log(1 - D(G(z)))] \tag{1}$$

where x is a real sample from the true data distribution and z is a random noise vector drawn from a distribution pz.

Our idea focuses on extracting facial expression information from identity representation. Thus, to learn face expression models, an encoder–decoder-based generator was employed to rebuild an expression image.

3.2 The proposed NST-GAN architecture

We consider the FER problems an image-to-image generation task where the goal is to generate facial images without identity given the input images. In our situation, we suppose that F_i^e and L_i^e represent an input and Synthetic facial image, respectively, where the superscript e denotes an expression, while the subscript i indicates an identity. Given an input facial image F_x^v of a subject x with expression v and a landmark variable L_z^y of a subject z with target expression y. The goal is to create a new free-identity facial image F' conditioned on the facial expression.

As shown in Fig. 2, our problem is described as follows:

$$F_x^v, L_z^y \rightarrow F_x'^y \tag{2}$$

In our model, the loss function for G and D of NST-GAN is defined as follows:

$$\min_D L_{\text{Disc}} = -ED(F_x^v) + ED(F_x'^y), \tag{3}$$

$$\min_G L_{\text{Gen}} = -EG(F_x'^y), \tag{4}$$

$$\min_{G,D} L_{\text{GAN}} = L_{\text{Disc}} + L'\text{Gen} \tag{5}$$

where $F_x'^y$ represents the generated image. The adversarial losses are optimized via WGAN-GP [24].

3.2.1 Generator

The generator runs an image-to-image as an auto-encoder task. For the most part, these models consist of a sequence of convoluted layers called encoders that perform sub-sampling up to a particular layer. Then, the operation is reversed into a sequence of oversampling and convolution layers called a decoder. There are two major drawbacks to using coder-decoder models for the proposed problem: First, the performance degrades quickly with the increase in the size of input images, and the model will have trouble retrieving lost information later. Second, the flow of image information crosses all layers, including the bottleneck. While inputs and outputs images share out some identical pixels, a massive amount of redundant features are exchanged. It is a waste of time and energy. Thus, we use the structure of a model called U-net [25] using skip connections. To improve the convergence and the performance of deep neural networks, skip connections are a widely used technique for training. For this reason, we added a Skip connection block every two

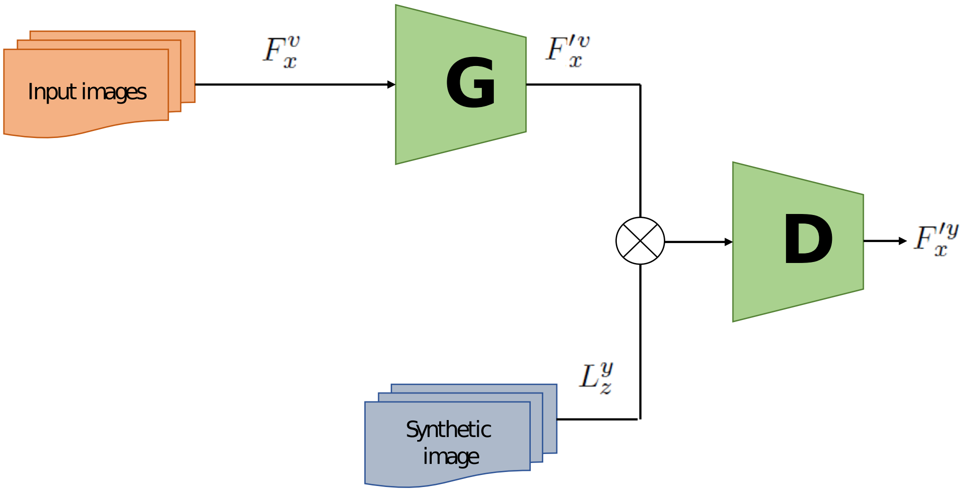

Fig. 2 The proposed approach of the NST-GAN illustrates two models: (1) a generator (G) which is a “U-Net” and (2) a discriminator (D)

layers to recuperate the image with less degradation. It is to note also that this technique is used to prevent the exploding problems and gradient vanishing. Batch normalization layers are evenly added to speed up the training. The generator architecture, used in this study, is illustrated in Fig. 3.

3.2.2 Discriminator

The proposed discriminator model is a simple Convolutional Neural Network (CNN), comprising five convolutional layers and fully connected layers for 6 classes. This model is shown in Fig. 4. The general expression of a convolution kernel is defined by Equation (6):

$$g(x, y) = \sum_{s=-1}^{a} \sum_{t=-b}^{b} W(s, t) f(x - s, y - t) \tag{6}$$

Where f is the original image, g is the filtered image and W is the filter kernel. Every segment of the filter kernel is inspected by $-a \le s \le a$ and $-b \le t \le b$. First, the discriminator takes into input two images: the faced image and its version (real or generated by the generator). The input images are concatenated in a $2 \times 256 \times 256$ shape tensor. Then, the new volume passes in the CNN model to end up in the last layer with a $1 \times 16 \times 16$ matrix. We employ ReLU as an activation function after each convolution layer. Finally, we use sigmoid as an activation function in the last layer. Furthermore, we append a fully connected layer to reassure that these nodes interact well. We apply dropout at the end of fully connected layers to inhibit our model from being oversized. Therefore, the final obtained matrix contains probabilities generated by the discriminator. After completing the training, the discriminator's function is finished. Given a facial image, we only use the generative network to remove identity information. But this discriminator must force the generator to produce better results during training. The image is equally transmitted to the model presented in our previous work [26] to read it and predict the expression.

4 Experiments and results

This section first introduces the experimental settings including the implementation details, the databases, and the preprocessing used in our experiment. Then, this section presents the experimental results of our proposed method and compares them with state-of-the-art. Finally, we evaluate the quality of the proposed approach using quantitative and qualitative evaluation to show the deep-learned feature's performance of each facial expression.

4.1 Experimental settings

We apply the proposed NST-GAN approach to the task of facial expression recognition on three publicly available facial expression databases: The Facial Expression Recognition 2013 (FER-2013) [27], The extended Cohn-Kanade (CK+) [18] and Japanese Female Facial Expression (JAFFE) [28].

4.1.1 Implementation details

The proposed approach was done on an Intel Core i7 PC, NVIDIA GeForce G920MX GPU, 16GB RAM, and the Ubuntu OS version LTS 18. Our architecture is implemented with Python 3, TensorFlow framework, and PyTorch deep learning framework. All experiments were built on GPUs only. GPU cores have the potential to accelerate processing. Thus, it reduces training times compared to a CPU.

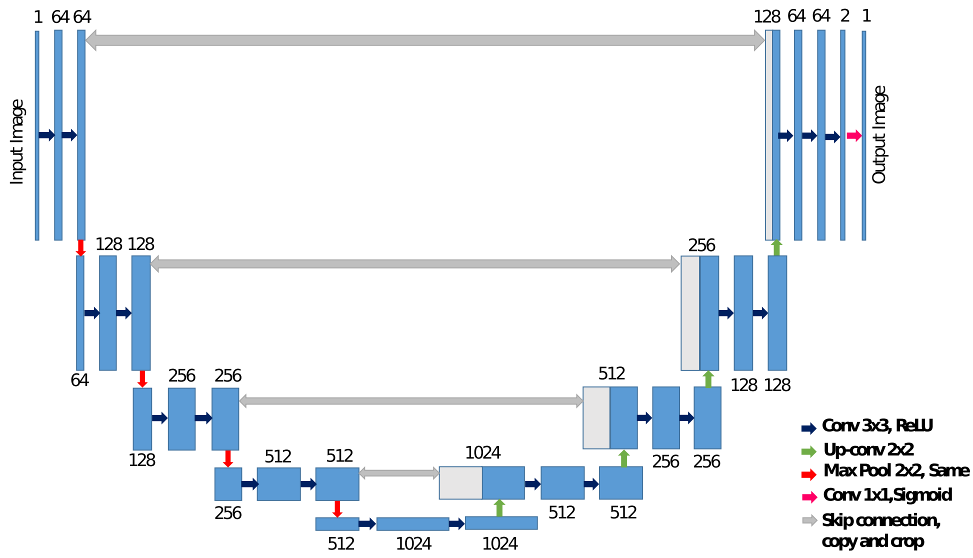

Fig. 3 Generator's architecture design used in this study

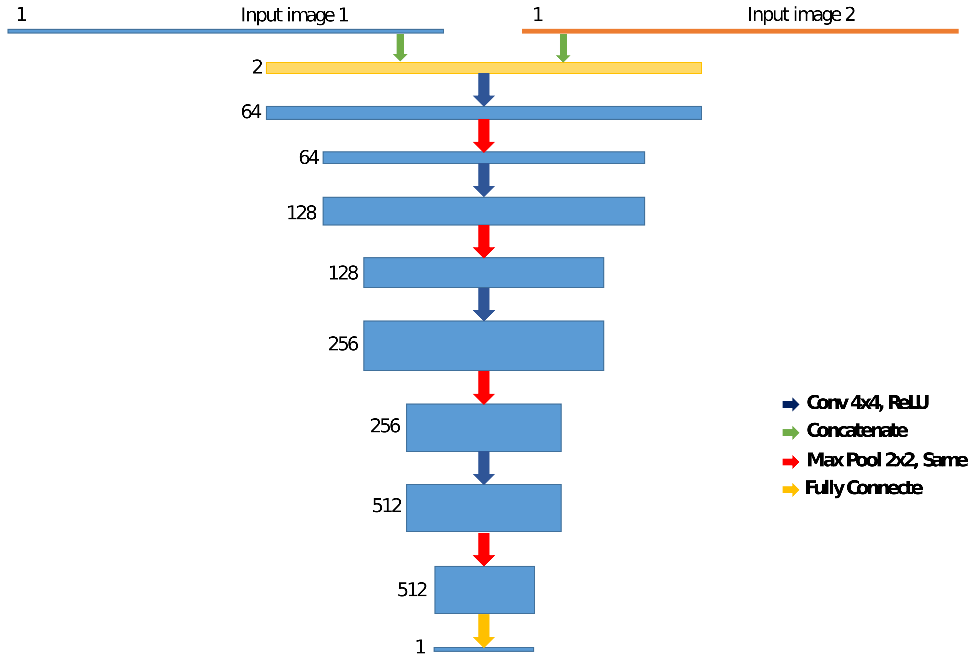

Fig. 4 Discriminator's architecture used in this study

4.1.2 Datasets

Due to the importance of human expression recognition in the design of HRI systems, several annotated databases have been created either in spontaneous uncontrolled settings such as FER-2013 [27] and AFEW [29] or in more strictly controlled environments as CK+ [18], MMI [30], and JAFFE [28]. Images captured in laboratory conditions (or controlled) are taken with a frontal pose, standard illumination, and common background conditions. However, we can consider this scenario as a solved problem. To illustrate the potency of the proposed NST-GAN, our experiments have been performed on three benchmark datasets, which are the Extended Cohn-Kanade (CK+), FER-2013, and JAFFE. Statistics describing these datasets are provided in Table 2.

- CK+ [18]: includes 593 videos recorded from 123 members aged in the middle of 18 and 30 years. The standard size of each frame is 640 x 490 pixels in PNG format. Besides, it presents 6 basic emotions. We chose only a tree frame from the last video sequence to collect samples of 6 basic expressions. We convert selected images to grayscale intensity images.
- FER-2013 [27]: is created by gathering the results of Google's Image Search API. It is a large-scale facial expression database with around 35,887 images in gray-

Table 2 Experiment set details for NST-GAN including the expression, training, and test samples

Expression	CK+		FER-2013		JAFFE	
	Tr. sam.[a]	Te. sam.[b]	Tr. sam.[a]	Te. sam.[b]	Tr. sam.[a]	Te. sam.[b]
Happiness	144	63	6292	2697	22	9
Sadness	60	24	4254	1823	22	9
Fear	52	22	3585	1536	23	10
Disgust	41	18	383	164	21	8
Anger	94	41	3467	1486	22	9
Surprise	174	75	2801	1201	22	9
Total	565	243	20,782	8907	132	54

[a] Training samples
[b] Test samples

scale values. All images are labeled as any of the seven emotions, and they are resized to 64 x 64 pixels.

- JAFFE [28]: includes 213 images exclusively in grayscale values. It presents 7 expressions of ten Japanese female models. The size of each face image is 256 x 256 pixels.

4.1.3 Preprocessing

Some conditions affect the expression recognition process and make this task a complex problem [26]. Major conditions include illumination, contrast, and size of input images. The pre-processing step detects the face via the Haar-cascade technique and reduces the lighting effects to some extent. To ameliorate the scaling variations, histogram equalization was employed to reduce the effect of illumination changes. In addition, to reduce the in-plane rotation, face alignment was utilized based on 3 facial key points, i.e., the tip of the nose and the centers of two eyes. For data augmentation purposes, we resized processed facial images to $N \times N$ with random rotation between $-4°$and $4°$and horizontal flipping.

4.2 Quantitative evaluation of the proposed approach

For all pre-training settings described in this paper, we conducted train our NST-GAN as follows:

- We fed the generator from the preprocessed images of size 64×64 as an input.
- The generated images are transmitted to the discriminator with the truth patches. Then, the discriminator begins to force the generator to produce outputs that are indistinguishable from the "real" images, while doing its best to detect the generator's "fakes".

The loss convergence would indicate that the model found some optimum, where it cannot improve more, which also should mean that it has learned well enough. If the loss has converged very well, it does necessarily mean that the model has learned very well as shown in Fig. 5. To validate the effectiveness of our approach, more procedures have been applied to estimate the skill of the proposed model on the three databases.

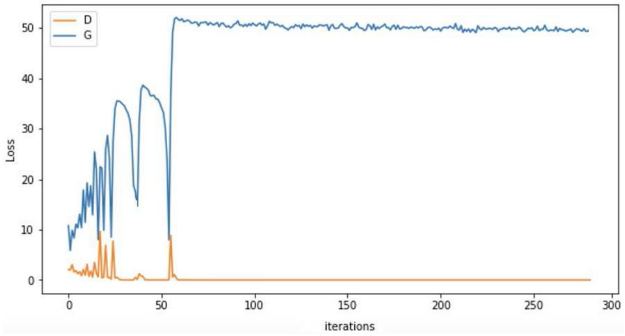

Fig. 5 Generator and discriminator loss during training

4.2.1 Subject-independent strategy

To gain more subtle data, we split each dataset into training, validation, and testing sets in a random subject-independent manner. On average, we attribute 60% for the training fold, 20% for the validation fold, and 20% for the testing fold. In each experiment, the proposed NST-GAN is trained for 1000 epochs.

As illustrated in Table 3, our approach obtains better or at least similar results compared to the state-of-the-art methods on all three datasets. For the performance of NST-GAN on the CK+ and FER-2013 database, the proposed model achieved the best recognition accuracy with 98.14% and 85.93%, respectively. Furthermore, the accuracy of the proposed model on JAFFE is similar to that of recent work. A comparison of the testing phase on three datasets is shown in Fig. 7. A considerable improvement is noticed from 400 epochs. Figure 6 presents an example of a sample built by the proposed NST-GAN generating identity-free expression from an input image of CK+ dataset.

Table 3 Proposed model versus other models performance comparison on CK+, JAFFE, and FER-2013

Method	CK+	JAFFE	FER-2013
Zhan et al. [31]	92.35	94.89	–
Khorrami [32]	–	82.43	–
Jain et al. [17]	–	94.91 8	–
Kumar et al. [18]	93.24	**95.23**	–
Lopes et al. [33]	92.73	94.86	–
Xie et al. [15]	97.59	–	72.14
Mollahossein et al. [34]	91.34	–	66.40
Proposed NST-GAN	**98.14**	95.12	**85.93**

Bold values indicate the databases used in this work

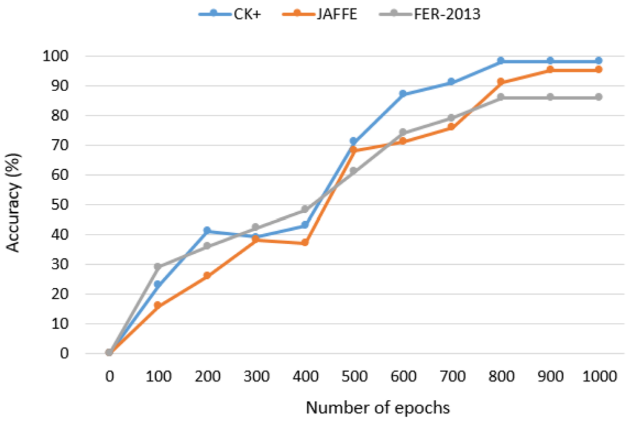

Fig. 7 Performance comparison on CK+, JAFFE, and FER-2013 databases per epocks

4.2.2 Per-class accuracy

Per-class accuracy (%) of CK+, JAFFE, and FER-2013 datasets is shown in Fig. 8 with the six expression classes that are happy, fear, surprise, disgust, sadness, and anger. It can be seen that neutral and happy classes are the top two with the highest accuracy rates in the three datasets. However, the accuracy of surprise and disgust expressions evaluated on the FER-2013 dataset is rather low compared with other classes. This decrease is due to the training sample numbers of surprise and disgust being much fewer than others shown in Table 2. Furthermore, compared to the CK+ database which performs high classification accuracy, the facial images in the JAFFE database are tougher to distinguish.

An example can be seen in Fig. 9; four samples (fear, surprise, sad, and disgust) from the FER-2013 dataset have only small differences and can be simply confused with each other. It leads to poor recognition performance. Overall, the record obtained in the classification accuracy evaluated on the FER-2013 dataset exhibits that the proposed NST-GAN is effective for expression classification in uncontrolled conditions.

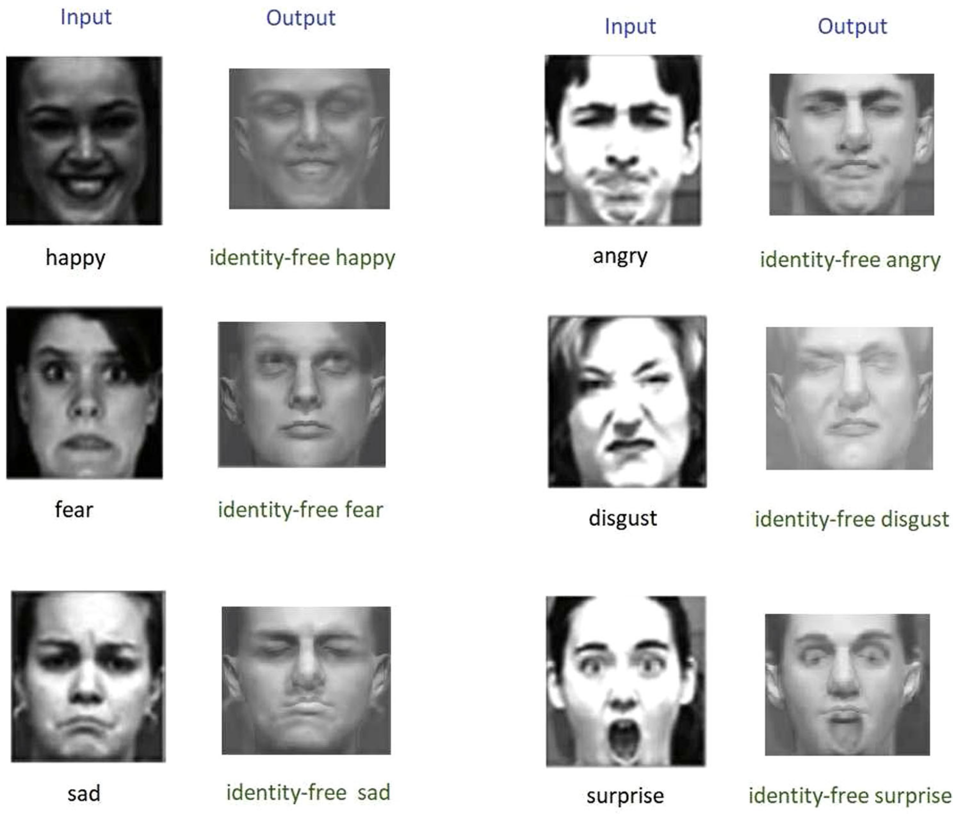

Fig. 6 Example of sample built by the proposed NST-GAN generating identity-free expression from an input image of CK+ dataset

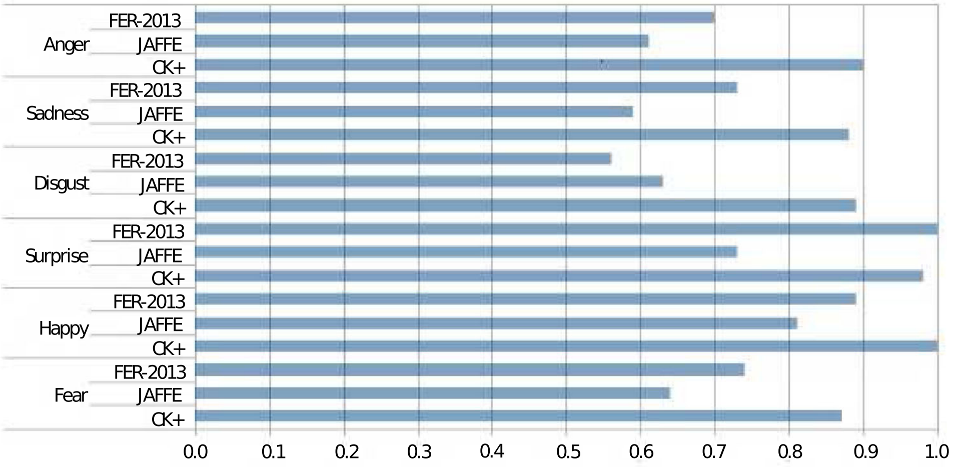

Fig. 8 Per-class accuracy (%) of CK+ dataset, JAFFE dataset, and FER2013 dataset with the six expression classes

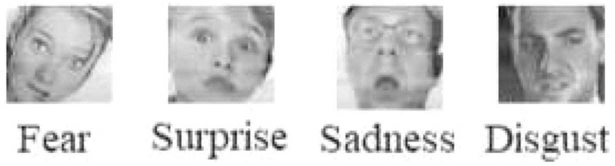

Fig. 9 Example of ambiguous facial samples of the four expressions (fear, surprise, sadness, and disgust) from the FER-2013 dataset

Table 4 Average accuracies of proposed NST-GAN in *K*-fold strategy on CK+, JAFFE, and FER-2013

K-fold	CK+	JAFFE	FER-2013
SET1 Test	0.96	0.74	0.86
SET2 Test	0.84	0.68	0.61
SET3 Test	1	0.94	0.98
SET4 Test	0.97	0.95	0.96
SET5 Test	0.89	0.87	0.79
SET6 Test	0.99	0.84	0.67
SET7 Test	1	0.87	0.95
SET8 Test	0.95	0.84	0.83
Average	**98.14**	**95.12**	**85.93**

Bold values indicate the databases used in this work

4.2.3 *K*-fold strategy

A *K*-fold strategy was employed eightfold for all three datasets. Each dataset was divided into 8 subsets. For each execution, we have specified 6 sets of data for training, and the rest subsets were assigned, respectively, for the validation and testing phases. As shown in Table 4, the reported results, in terms of accuracy, are the average of the 8 executions in the testing phase.

Table 5 Experiment set details for NST-GAN including the expression, training, and testing samples

	Testing		
Training	CK+	JAFFE	FER-2013
CK+	–	63.69	52.30
JAFFE	51.35	–	**57.90**
FER-2013	**85.26**	**76.58**	–

Bold values indicate the databases used in this work

4.2.4 Cross-database strategy

The cross-database strategy (training and testing) was performed to test the generalization performance of our proposed approach. Table 5 presents the accuracy of each database when one of the three is used for the training phase and the remaining ones are used for the testing phase. We noticed that the accuracies of FER-2013 trained using other databases are relatively low because the images are all collected under controlled conditions. The model trained using JAFFE achieved only an average rate of accuracy on FER2013 and CK+ databases. The highest accuracy (85.26%) is obtained for the CK+ database when the proposed model was trained using the FER-2013 database.

4.2.5 Expression-specific performance results

Besides cross-validation evaluation, one of the most significant indicators to evaluate a model is confusion matrices. Its performance based on right and wrong predictions is revealed and broken down by class. Sequel to this fact, three confusion matrices have been computed on the CK+, JAFFE, and FER-2013 databases and sit in Tables 6, 7 and 8, respectively. Confusion matrices are performed where the proposed model has the top cross-validation performance.

Table 6 Confusion matrix with NST-GAN on CK+ database

	Happiness	Sadness	Fear	Disgust	Anger	Surprise
Happiness	**93.0**	0.0	0.0	0.0	7.0	0.0
Sadness	0.0	**100**	0.0	0.0	0.0	0.0
Fear	0.0	0.0	**91.2**	6.5	0.0	2.3
Disgust	0.0	3.9	0.0	**95.3**	0.8	0.0
Anger	0.0	6.4	0.0	0.0	**89.5**	4.1
Surprise	0.0	0.0	0.0	0.0	0.0	**100**

Bold values indicate the databases used in this work

Table 7 Confusion matrix with NST-GAN on JAFFE database

	Happiness	Sadness	Fear	Disgust	Anger	Surprise
Happiness	**92.3**	2.5	0.0	5.1	0.1	0.0
Sadness	0.0	**94.0**	0.0	6.0	0.0	0.0
Fear	0.0	11.0	**89.0**	0.0	0.0	0.0
Disgust	0.0	5.0	0.0	**88.6**	0.0	6.4
Anger	0.0	0.0	0.0	20.0	**69.8**	10.2
Surprise	0.0	0.0	9.0	11.5	20.9	**58.6**

Bold values indicate the databases used in this work

Table 8 Confusion matrix with NST-GAN on FER-2013 database

	Happiness	Sadness	Fear	Disgust	Anger	Surprise
Happiness	**91.0**	8.0	0.0	0.0	0.0	1.0
Sadness	0.0	**83.3**	0.0	0.0	0.0	16.7
Fear	0.0	0.0	**69.0**	21.8	9.2	0.0
Disgust	0.0	0.0	5.6	**70.7**	0.0	24.3
Anger	0.0	0.0	14.0	0.0	**86.0**	0.0
Surprise	0.0	0.0	10.0	0.0	0.0	**90.0**

Bold values indicate the databases used in this work

4.3 Qualitative evaluation of the proposed approach

To view the deep-learned features of each facial expression, T-distributed Stochastic Neighbor Embedding (t-SNE) [35] is applied. The T-SNE is a nonlinear algorithm allowing to show learned large-dimensional features in a three-dimensional space. This algorithm is based on a probabilistic interpretation of proximities. A probability distribution is defined over pairs of points in the original space such that points close to each other have a high probability of being chosen while points far apart have a low probability of 'being selected. Figure 10 shows a 3D t-SNE plot of the deep-learned features from the FER-2013 dataset. The output of the last fully connected layer of discriminator D presents the deep-learned features. The random sample number is fixed to 1150 considering the computing speed and the validation set number of the FER-2013 dataset. As observed in the 3D t-SNE plot, the dots of disgust expression are relatively few due to the unbalanced data distribution of the training set. As it can be seen in Table 2, the available training samples number of disgust in the FER-2013 database is only 383, while the training number of other expressions surpassed 2800. Although various information exists in the database (including identity information, age, race, etc.), the dots of each expression are distributed and there is a perfectly clear interval between seven expressions which demonstrates the effectiveness of our NST-GAN model.

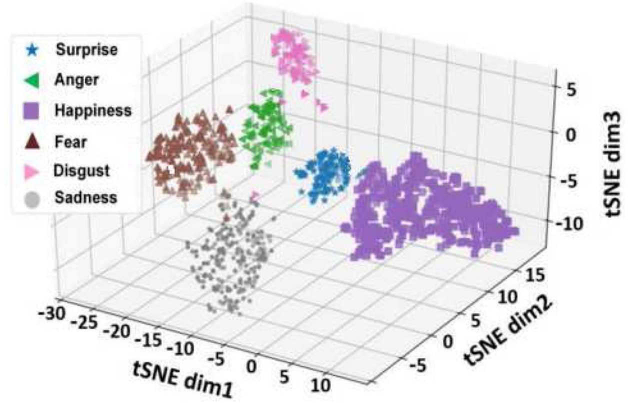

Fig. 10 The 3D t-SNE visualization by the proposed model on the FER-2013 dataset

5 Conclusion

In this work, we propose an NST-GAN-based approach to removing the facial identity factor from facial images and training the generative and discriminative representations simultaneously from synthetic identity. The disentangling task of the facial identity factor is performed in two phases: learning by a conditional generator G and learning by a discriminator D. We have conducted qualitative and quantitative experiments on the proposed method, which exhibit that our NST-GAN outperformed recent works when all the judging criteria were taken. The experimental results prove that the NST-GAN is effective for expression recognition and exceeds recent state-of-the-art systems on three well-known facial expression datasets, i.e., CK+, JAFFE, and FER-2013. The cross-validation strategy also demonstrates the promising generalization potential of our method. One limitation of this work is that the model has not intrinsic metric evaluation present for better model training and generating complex outputs. Thus, we still cannot evaluate the generative models. Recognizing other emotional symptoms are worth investigating in this domain where vocal and gestures have a relative role to recognize human emotion. In other words, an intelligent system should be able to rightly recognize social signals that interpret.

Acknowledgements Faten Khemakhem and Hela Ltifi declare that they have no affiliations with or involvement in any organization or entity with any financial interest, or non-financial interest in the subject matter or materials discussed in this manuscript.

Author contributions All authors contributed equally to this manuscript.

Data availability The CK+ dataset is included in this published article [18]. The FER-2013 dataset is included in [27]. The JAFFE dataset is included in [28].

Declarations

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

References

1. Ekman P, Friesen WV (1971) Constants across cultures in the face and emotion. J Pers Soc Psychol 17(2):124. https://doi.org/10.1037/h0030377
2. Martinez B, Valstar MF (2016) Advances, challenges, and opportunities in automatic facial expression recognition. Springer, Cham, pp 63–100. https://doi.org/10.1007/978-3-319-25958-1_4
3. Christopher P, Martin K (2016) Facial expression recognition using convolutional neural networks: state of the art. CoRR arXiv:1612.02903
4. Zhang X, Mahoor MH, Mavadati SM (2015) Facial expression recognition using l_p-norm MKL multiclass-SVM. Mach Vis Appl 26:467–483. https://doi.org/10.1007/s00138-015-0677-y
5. Liu Z, Wu M, Cao W, Chen L, Xu J, Zhang R, Zhou M, Mao J (2017) A facial expression emotion recognition based human–robot interaction system. IEEE/CAA J Automatica Sinica 4(4):668–676. https://doi.org/10.1109/JAS.2017.7510622
6. Tao L, Matuszewski BJ (2016) Is 2D unlabeled data adequate for recognizing facial expressions? IEEE Intell Syst 31(3):19–29. https://doi.org/10.1109/MIS.2016.25
7. Baltrusaitis T, Robinson P, Morency L-P (2016) Openface: an open source facial behavior analysis toolkit. In: 2016 IEEE winter conference on applications of computer vision (WACV), pp 1–10. https://doi.org/10.1109/WACV.2016.7477553
8. Luo Y, Wu C-M, Zhang Y (2013) Facial expression recognition based on fusion feature of PCA and LBP with SVM. Optik Int J Light Electron Opt 124(17):2767–2770. https://doi.org/10.1016/j.ijleo.2012.08.040
9. Chen J, Chen Z, Chi Z, Fu H (2014) Facial expression recognition based on facial components detection and hog features
10. Dhall A, Goecke R, Joshi J, Sikka K, Gedeon T (2014) Emotion recognition in the wild challenge 2014: Baseline, data and protocol. In: Proceedings of the 16th international conference on multimodal interaction. ICMI '14, pp 461–466. Association for Computing Machinery, New York. https://doi.org/10.1145/2663204.2666275
11. Hertel L, Barth E, Käster T, Martinetz T (2015) Deep convolutional neural networks as generic feature extractors. In: 2015 International joint conference on neural networks (IJCNN), pp 1–4 . https://doi.org/10.1109/IJCNN.2015.7280683
12. Han D, Liu Q, Fan W (2018) A new image classification method using CNN transfer learning and web data augmentation. Expert Syst Appl 95:43–56. https://doi.org/10.1016/j.eswa.2017.11.028
13. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–444. https://doi.org/10.1038/nature14539
14. Jiuxiang G, Zhenhua W, Jason K, Lianyang M, Amir S, Bing S, Ting L, Xingxing W, Gang W, Jianfei C, Tsuhan C (2018) Recent advances in convolutional neural networks. Pattern Recogn 77:354–377. https://doi.org/10.1016/j.patcog.2017.10.013
15. Xie W, Jia X, Shen L, Yang M (2019) Sparse deep feature learning for facial expression recognition. Pattern Recogn 96:106966. https://doi.org/10.1016/j.patcog.2019.106966
16. Kaya M, Bilge HS (2019) Deep metric learning: a survey. Symmetry. https://doi.org/10.3390/sym11091066
17. Jain N, Kumar S, Kumar A, Shamsolmoali P, Zareapoor M (2018) Hybrid deep neural networks for face emotion recognition. Pattern Recognit Lett 115:101–106. https://doi.org/10.1016/j.patrec.2018.04.010. (**Multimodal fusion for pattern recognition**)
18. Jain DK, Pourya S, Paramjit S (2019) Extended deep neural network for facial emotion recognition. Pattern Recognit Lett 120:69–74. https://doi.org/10.1016/j.patrec.2019.01.008
19. Yi Z, Zhang H, Tan P, Gong M (2018) DualGAN: unsupervised dual learning for image-to-image translation
20. Souibgui MA, Kessentini Y (2022) De-gan: a conditional generative adversarial network for document enhancement. IEEE Trans Pattern Anal Mach Intell 44(3):1180–1191. https://doi.org/10.1109/tpami.2020.3022406
21. Zhang F, Zhang T, Mao Q, Xu C (2018) Joint pose and expression modeling for facial expression recognition. In: 2018 IEEE/CVF conference on computer vision and pattern recognition, pp 3359–3368. https://doi.org/10.1109/CVPR.2018.00354
22. Yang H, Ciftci U, Yin L (2018) Facial expression recognition by de-expression residue learning. In: 2018 IEEE/CVF conference on computer vision and pattern recognition, pp 2168–2177. https://doi.org/10.1109/CVPR.2018.00231
23. Goodfellow I.J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial networks, vol 27. arXiv:1406.2661
24. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville A (2017) Improved training of wasserstein gans. In: Proceedings of the 31st international conference on neural information process-

ing systems. NIPS'17, pp 5769–5779. Curran Associates Inc., Red Hook

25. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. CoRR arXiv:1505.04597
26. Khemakhem F, Ltifi H (2019) Facial expression recognition using convolution neural network enhancing with pre-processing stages. In: 2019 IEEE/ACS 16th international conference on computer systems and applications (AICCSA), pp 1–7. https://doi.org/10.1109/AICCSA47632.2019.9035249
27. Goodfellow IJ, Erhan D, Carrier PL, Courville A, Mirza M, Hamner B, Bengio Y (2013) Challenges in representation learning: a report on three machine learning contests. In: ICONIP, vol 8228. https://doi.org/10.1007/978-3-642-42051-1_16
28. Lyons M, Akamatsu S, Kamachi M, Gyoba J (1998) Coding facial expressions with Gabor wavelets. In: Proceedings third IEEE international conference on automatic face and gesture recognition, pp 200–205. https://doi.org/10.1109/AFGR.1998.670949
29. Dhall A, Goecke R, Lucey S, Gedeon T (2012) Collecting large, richly annotated facial-expression databases from movies. IEEE Multimedia 19(3):34–41. https://doi.org/10.1109/MMUL.2012.26
30. Pantic M, Valstar M, Rademaker R, Maat L (2005) Web-based database for facial expression analysis. In: 2005 IEEE international conference on multimedia and expo, p 5. https://doi.org/10.1109/ICME.2005.1521424
31. Zhang T, Zheng W, Cui Z, Zong Y, Li Y (2019) Spatial–temporal recurrent neural network for emotion recognition. IEEE Trans Cybern 49(3):839–847. https://doi.org/10.1109/TCYB.2017.2788081
32. Khorrami P, Le Paine T, Brady K, Dagli C, Huang TS (2016) How deep neural networks can improve emotion recognition on video data. In: 2016 IEEE international conference on image processing (ICIP), pp 619–623. https://doi.org/10.1109/ICIP.2016.7532431
33. André TL, Edilson D, Alberto FD, Thiago O-S (2017) Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. Pattern Recognit 61:610–628. https://doi.org/10.1016/j.patcog.2016.07.026
34. Ali M, David C, Mahoor MH (2016) Going deeper in facial expression recognition using deep neural networks. In: 2016 IEEE winter conference on applications of computer vision (WACV), pp 1–10
35. Laurens VDM, Geoffrey H (2008) Visualizing data using t-SNE. J Mach Learn Res 9(86):2579–2605

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.