Generative Adversarial Networks for Facial Expression Recognition in the Wild

Luma Alharbawee[1] and Nicolas Pugeault[2]

Abstract: The task of modeling and identifying people's emotions using facial cues is a complex problem in computer vision. Normally we approach these issues by identifying Action Units, which have many applications in Human Computer Interaction. Although Deep Learning approaches have demonstrated a high level of performance in recognizing AUs and emotions, they require large datasets of expert-labelled examples. In this article, we demonstrate that good deep features can be learnt in an unsupervised fashion using Deep Convolutional Generative Adversarial Networks, allowing for a supervised classifier to be learned from a smaller labelled dataset. The paper primarily focuses on two key aspects: firstly, the generation of facial expression images across a wide range of poses (including frontal, multi-view, and unconstrained environments), and secondly, the analysis and classification of emotion categories and Action Units. Utilizing a pioneering methodology and incorporating an extensive array of datasets for feature acquisition and classification, we substantiate a remarkably persuasive generalization and achieve enhanced outcomes. In contrast to prevailing state-of-the-art techniques, our proposed model showcases exceptional performance, specifically on the Radboud dataset, boasting an unparalleled overall accuracy rate of 98.57%.

Keywords: Affective computing; GANs; DCGAN; fine-tuning; transfer learning; relabelling; generalisation; FACS.

1. INTRODUCTION

Emotion recognition presents a significant and complex challenge within the realm of computer vision. Particularly, its integration into Human-Computer Interaction holds exceptional significance. The challenge in modelling and identifying emotions emerges when individuals exhibit significant variations in facial features, alongside the extensive range of expressions observed across diverse individuals, cultures, and contexts. Emotions are commonly delineated utilizing individual AUs, which serve as the fundamental building blocks of facial expressions related to emotions.

The advent of advanced Deep Learning (DL) approaches in recent times has yielded remarkable advancements in the realm of automatic facial emotion class recognition. Nevertheless, DNNs necessitate a significant volume of training data. By employing a substantial training set, the problem of overfitting is mitigated, facilitating enhanced generalization and acquisition of superior features. Furthermore, a larger training set proves to be more efficient in comprehending intricate relationships and patterns that exist within the data distribution. Nonetheless, accessing or having a dataset that has a level of labelled coverage that is sufficient across several situations and conditions is a comparatively large challenge. In addition, effectively producing authentic and dynamic facial expressions that accurately reflect facial AUs remains a formidable challenge, primarily due to the continued difficulty in automatically recognizing the intensity of AUs [1].

Facial expression datasets with Action Units and emotion labels are scarce, limited in size, and imbalanced [2] due to the scarcity in the diversity of certain emotions and AUs. Furthermore, the process of labelling facial expressions is challenging, requiring significant effort, cost, time, and expertise [3]. In certain domains such as remote sensing, qualified experts are typically needed to perform this task since publicly available satellite images and their corresponding ground truth data are often not provided. The result is that there is not enough data to optimise all parameters, yet the quantity of labelled data is rarely enough to constrain numerous parameters. The consequence is that the models become prone to overfitting and demonstrate an inadequate ability to generalize when exposed to unseen subjects, as documented by Han et al. (2016). Research has conclusively indicated that the effectiveness of Deep Learning in generalizing improves proportionally with the inclusion of a substantial amount of nonlinear facial variability factors in the training data. These factors, which comprise individual distinctions, subject identity, facial morphology, various backgrounds, illuminations, occlusions, and head pose, are frequently encountered in unconstrained environments [4]. Hence, a considerable amount of research in the field of Deep Learning has been dedicated to various aspects such as maintaining balanced batches, utilizing ReLU activation functions, training on multiple datasets, implementing dropout regularisation, harnessing GPU acceleration [5], and employing data augmentation techniques. The main objective of these augmentation techniques is to increase the size of the training data in order to better

E-mail address: Luma.akram@uomosul.edu.iq, Nicolas.Pugeault@glasgow.ac.uk

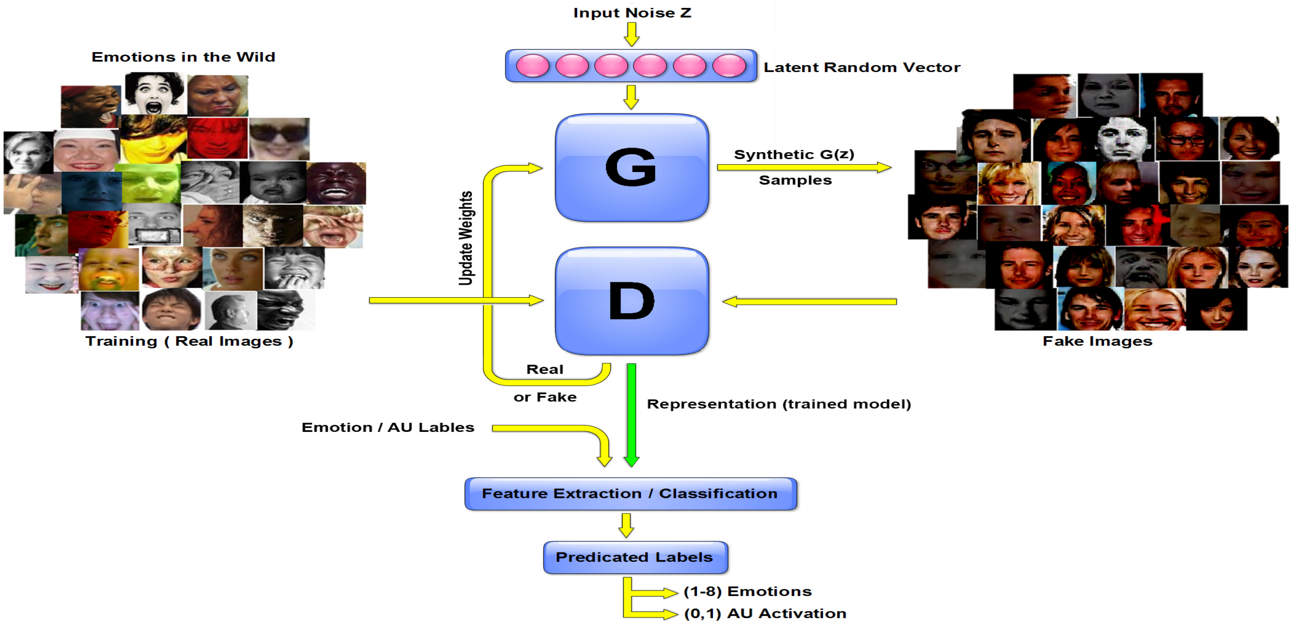


Figure 1. This figure showcases the structure of a Deep Convolutional GAN (DCGAN), which integrates the fundamental elements from Convolutional Neural Networks (CNNs) used in supervised learning alongside the conventional principles of GANs employed in unsupervised learning models. The proposed model involves the utilization of two deep neural network models (G & D), namely G and D, wherein G denotes the Generator and D signifies the Discriminator. The primary aim of the Generator is to generate synthetic images with a high degree of resemblance to genuine ones, thereby effectively deceiving the Discriminator. The discriminator works as a CNN-based classification network and its output class probabilities. Both models are trained jointly in a competitive min-max process. The process achieves a state of equilibrium when the Discriminator no longer discerns between genuine and counterfeit images. The latter part of this architecture comprises the stages of extracting features and classifying facial expressions. Once the features are extracted from the DCGAN discriminator, they are used as input for a Support Vector Machine (SVM) model to facilitate the final classification and detection. The images that have been used for training the SVM are the initial ones from each database, and that differs from experiment to experiment depending on the dataset used for training and testing.

represent the actual distribution of the problem domain. This leads to a broader range of variations and diversity in the dataset. The various methods mentioned above contribute to improving the quality performance of deep NNs and increasing the quantity of the dataset. However, a limitation is that they do not sufficiently address the requirement for non-linear parametric variations in training datasets, which may not be addressed by traditional augmentation approaches. In light of this, an alternative option is to utilize a sizeable unlabelled dataset and employ unsupervised learning methods. Although there is an increasing amount of available data from the internet, most are unannotated. Therefore, one way to exploit the available unlabelled data, and give an incentive to use unsupervised learning, is to learn better representations that can be used with these supervised tasks. Meanwhile, technically, a solution to alleviate these obstacles is to innovate data models using synthetic data accompanied by genuine data to train these models. DCGAN, which stands for Deep Convolutional Generative Adversarial Network, is an advanced technique used to generate facial images. This method has gained significant popularity due to its exceptional performance in creating realistic and high-quality facial images [6]. This method provides a balanced approach to tackle a wide range of computer vision problems. These problems include modifying facial attributes, exploring reinforcement learning, translating synthetic images into realistic photos, synthesizing images in different styles, transforming images, restoring colours, creating textures, augmenting datasets, generating shop advertisements [7], analyzing sentiments [3], translating images, editing face generation, editing human poses [8], processing natural language, colourizing images, and adjusting poses [9].

Figure 1 gives an overview of the proposed approach. The model implements two Deep Neural Networks G & D, where G is the *Generator* and D is the *Discriminator*, as is typical for GANs. The Generator is trained to generate fake images similar enough to the real ones to fool the Discriminator. Both models are trained in a competitive min-max process at the same rate, on an unlabelled large dataset of facial images. The persistent confrontation training among the generator structure and the discriminator structure would improve both the discriminator's identification ability and the accurate extraction of image features. These automatic features engineering or representation learning are suggested to indicate that the input comes from the training dataset [10]. In this context,

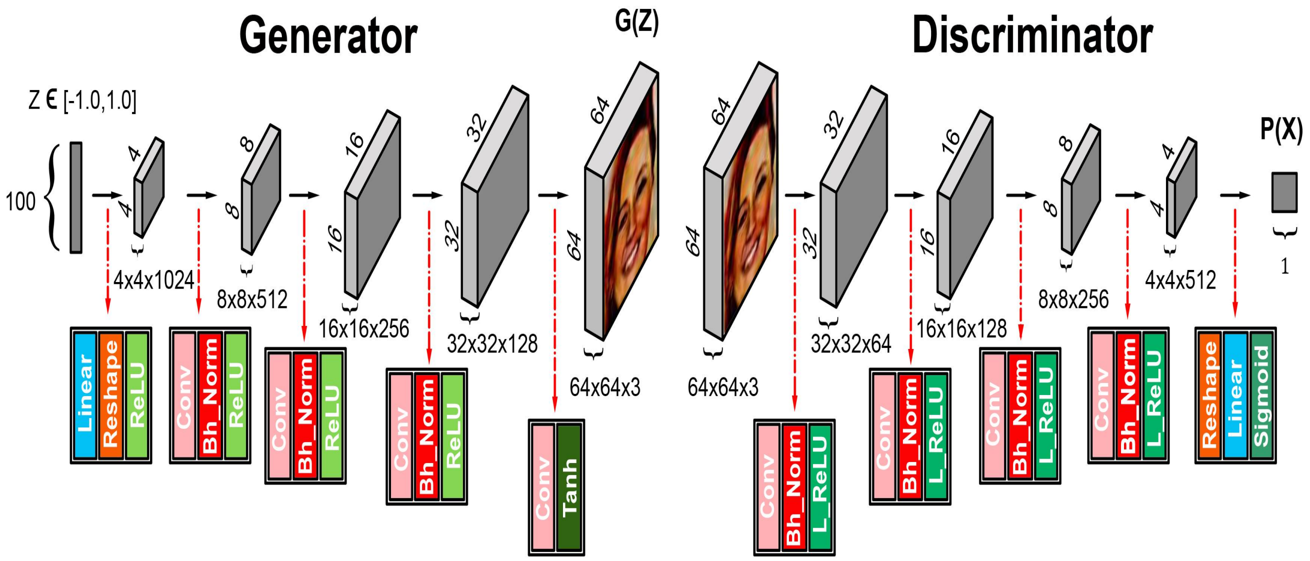


Figure 2. A sketch of DCGAN architecture.

the adversarial training process is repeated until the Nash equilibrium [11] is reached between the Generator and the Discriminator to achieve good images. The traditional DCGAN model is trained with an aggregation of log loss on the Discriminator output and ℓ_1 loss between the Generator output and target image. The Discriminator is only trained with log loss. To interpret the loss when training DCGAN, the Discriminator and Generator would adjust their weights with the value function in the equation below. The objective requires the Generator to produce data that can match the statistics of the real data. In this case, the Discriminator is only used to match whichever statistics are identical. The G and D sub-network's minimax objective function can be optimized during training by adjusting the loss function:

$$\min_G \max_D V(D,G) = Ex \sim pdata(x)[\log(D(x))] + Ez \sim pz(z)[\log(1-D(G(z)))]$$

Following the initial pre-training phase, the Discriminator network that has been trained is subsequently employed as a tool for extracting relevant features. These extracted features are further utilized in training SVM classifiers to identify emotions and AUs, using a much smaller labelled dataset. Finally, the trained model will be deployed for the supervised task for the classification of facial expressions with the available emotion and AU labels. Figure 2 represents the DCGAN architecture chart.

A question is whether it is possible to enhance the categorization of emotions captured in uncontrolled settings by employing Generative Adversarial Networks, specifically DCGAN. Can the features extracted from the Discriminator be utilized for successful facial action unit (AU) and emotion recognition? Is there a method to achieve consistent generalization across diverse datasets?

The key findings and achievements of this study can be outlined in the following manner:

- Utilizing unsupervised Generative Adversarial network models effectively as a feature extraction method in supervised tasks for recognizing facial expressions in uncontrolled environments. It examines the application of DCGAN in extracting facial characteristics and classifying seven emotions in natural settings, along with Action Units. A constructive framework was proposed by using the Discriminator network as a feature extractor based on video frames and static images. More precisely, testifying was done to see whether the features learned from the Discriminator's convolutional penultimate layer could provide information characterizing emotions and AUs.
- The ability of DCGAN to generate arbitrary analogous images from a different perspective (predefined in front, multi-view settings and from real-life wild conditions) was discovered, which was indiscernible from their versions in unsupervised manner adaptation. A set of four quantitative metrics, namely Inception Score (IS), Fréchet Inception Distance (FID), Structural Similarity Index (SSIM), and the Amazon Mechanical Turk (AMT), were employed to assess the quality of the generated samples across all the datasets. Furthermore, an assessment was conducted to examine if the generated samples exhibited any

signs of mode collapse. Moreover, a thorough discussion was provided concerning these observations.

- A manual re-annotation of the images of the Radboud dataset (emotions relabelled to AUs) was achieved. Higher quality discriminative representation features were derived from a large number of examples and from frontal face images.
- A generalisation across datasets' evaluation performances was presented, using various pre-trained models to cope with the impact of the restricted number of the target dataset. Additionally, we examined how the features trained on a large dataset that is potentially unlabelled can be experimentally transferred from the supervised task to a different one.

This work is unique in that it formulates the usage of unsupervised Generative Adversarial Networks models as a feature extraction for the supervised tasks. This is for facial features' extraction and for classifying the seven emotion classes *(Fearful, Disgusted, Angry, Sad, Happy, Contemptuous, and Surprised)* in the wild together with Action Units. Modeling and accurately discerning individuals' emotions solely based on their facial expressions is a complex and challenging problem within the realm of computer vision. Typically, emotions have been described and categorized by identifying specific Action Units (AUs), which are the fundamental units comprising facial expressions. In this context, the proposed system not only builds upon existing research findings but also pushes the boundaries of current knowledge by investigating how emotional cues can be effectively learned and recognized through the exploration of subtle local changes in facial appearance. Furthermore, the system addresses the crucial aspect of generalization by studying how these learned patterns can be extrapolated and effectively applied to new individuals. This approach stands to contribute significantly to the advancement of emotion recognition technologies [12]. Effective facial expression recognition is crucial in a wide range of practical applications. It offers numerous benefits in fields like advanced human-computer interaction, robotic systems, affective computing, security, machine learning, stress, and depression analysis. Consequently, machines interacting with people need to possess reliable facial expression recognition capabilities to effectively meet the diverse demands of these applications [13].

2. Related work

Considerable effort has been dedicated to synthesizing images with the aid of generative deep learning approaches. Early work comprises of Restricted Boltzmann Machines and includes their variants such as Deep Belief Networks [14]. More recently, various successful models have been developed in the area, including the Auto-Regressive models [15], and the Variational Auto Encoders. Despite the robust and stable training of Variational Autoencoders (VAEs) [16], they tend to generate images with blurred details [17]. Conversely, when VAEs and Generative Adversarial Networks (GANs) are jointly trained [18], it becomes feasible to utilize the distinctive feature representations acquired by the discriminator of a GAN in order to improve the objective of VAE reconstruction. It also enables the learning of an identity-invariant information representation [19]. Recently new approaches have used GANs [20]. DCGAN is known to have a higher level of performance in image generation [21]. DCGAN merges GAN and CNN to provide techniques for enhancing the training stability [22]. More recent work uses conditional GAN and added auxiliary constraints for augmenting the model. This uses class labels and, for governing the generator output and the discriminator, was exploited as a classifier to predict the classes. Luan et al. [23] and Springenberg [24] generalized GAN by learning a discriminative classifier, where D was trained not just to differentiate between real and non-genuine subjects, but also to classify the processed images [25]. Zheng [26] also concentrates on semi-supervised training by allocating a unified label distribution upon all the current classes of GAN samples [22]. The proficiency of GANs in image processing is exemplified by a range of other GAN applications, including but not limited to the generation of frontal facial images from rotated ones [27], the alteration of images to preserve the identity of the depicted subject, and the elimination of excessive illumination in facial images to create optimal conditions for facial identification [28]. These instances serve as additional evidence to validate the efficacy of GANs in the realm of image processing. The classifiers yield multiple outputs due to the various training methodologies available for GANs, such as supervised learning, unsupervised learning, and semi-supervised learning [28].

3. Methodology

Ian Goodfellow et al. first introduced the idea of Generative Adversarial networks (GANs) [29]. To enhance the training stability and performance of GANs, the DCGAN framework was developed. This framework has demonstrated its stability and power by generating synthetic images that closely resemble real images. This work is extensively described in the following sections, which outline the main steps in more details: The ability of the DCGAN model was adapted for supervised tasks by deep facial features, which were extracted and grounded on this model. After training the model, it was observed whether the generated images of AUs and emotions have the same visual fidelity quality of the original images. In terms of assessing their generalisation ability, the trained models were validated on more datasets: RPI ISL Enhanced Cohn-Kanade [30], Large-scale CelebFaces Attributes(CelebA) [31], Radboud Faces Database(RaFD) [32], Real-world Affective Faces Database(RAF-DB) [33], Karolinska Directed Emotional Faces(KDEF) [34], and Static Facial Expressions in the Wild(SFEW) [35] using the transfer learning approach. Table 1 summarises the datasets of emotions and AUs used in this work. The Viola-Jones method suggested by [36] was used to crop frontal images. Additionally, the

MTCNN [37] approach, which is a state-of-the-art, multi-task CNN method, was utilized to obtain cropped faces from multiple viewpoints. This method was employed for both facial landmark recognition and bounding box delineation. The images were then downscaled to an initial resolution of 64 × 64 pixels before being inputted into the network. The model was then trained for a span of 300 epochs. The features from the Discriminator's convolutional penultimate layer 12 were extracted; this layer gives 512 feature spatial grid maps of size 64 × 64. Then, the singleton dimensions of size 1 were reshaped and removed from the shape of a tensor (4-dimensional tensor).

The nonlinear SVM was used for emotion recognition and the linear SVM was used for AUs activation detection, alongside the emotion/AU labels. SVM was straightforwardly applied at the top of these features to predict and recognise the occurrence of 14 AUs and eight emotion classes as training end to end. The same steps could be used to extract the features from the Generator, but this can be a task for future research.

From a technical standpoint, DCGAN code execution prerequisites are required training on the GPU computing capabilities that necessitate the Parallel Computing Toolbox and a CUDA (Graphics card) implementation of enabled NVIDIAGPU. We performed our framework using the Deep Learning Toolbox 2022a MATLAB implementations of Deep Convolutional Generative Adversarial Networks (DCGANs). The execution of the DCGAN model utilised the MatConvNet library. MatConvNet (CNNs using MATLAB) is an efficient MATLAB toolbox implementation of the Convolutional Neural Networks (CNNs) models for the applications of computer vision. It can run, learn and implement most state-of-the-art CNN algorithms.

TensorFlow represents a development tool for second-generation flexible arrangement for both the Google company and the deployment of numerous machine learning applications. It can be used to create neural networks. This setup was made by using the compatible integration of the Tensor Flow and CUDA toolkit to empower the parallel calculation and allow better computation execution times and performance. The experiments were executed on the workstation, specifically employing the Ubuntu Linux system. To expedite the training and testing processes, the utilization of NVIDIA GeForce GTX 980 Ti GPUs was incorporated.

The model was then trained using the following hyperparameter values: the optimization algorithm utilized a mini-batch SGD with a batch size of 128. The learning rate for the optimizers was fixed at 0.0002, and a momentum coefficient term, denoted by β_1, was chosen to be 0.9 to enhance training stability. Furthermore, the Adam optimizer (Adaptive Moment Estimation) was adopted as the most suitable choice for minimizing the loss function based on extensive research in the field of generative models. The weights were initialized using a zero-centered normal distribution with a standard deviation of 0.02. To ensure input normalization for each unit, batch normalization was employed, resulting in a standardized distribution characterized by zero mean with variance. We depended on using the DCGAN architecture with the available adjusted hyper-parameter values as described in their design. These hyper-parameters have been recommended for the training of the model. Additionally, cross-validation was conducted to discover the best hyper-parameters and assess the model's performance with the highest accuracy. Figure 3 illustrates the examination of Generator and Discriminator loss for each batch during the training of the DCGAN, pertaining to all datasets utilized in this research. Figure 4 visualizes all the generated images for all the datases used in this work.

4. Experiments

Two separate series of experiments were conducted to evaluate the proposed approach for both emotion recognition (section 4-A) and AU recognition (section 4-B).

A. *Experiments on Emotion Recognition*

The primary objective of this experiment is to accurately identify and classify emotions, as well as produce corresponding facial expression images. This was briefly divided into eight experiments, to show that the performance gained whether dependent on the specific dataset or was provided from different datasets. Training a new DCGAN involves utilizing the generative model's capability to produce diverse images with varying perspectives, including frontal, multi-view, and in real-life scenarios. The concept of cross dataset evaluation was established by considering various datasets and the fundamental principles of transfer learning and different pre-training models. This was subsequently elucidated as outlined below:

1) Testing a pre-trained model of the enhanced CK dataset (source dataset) on the frontal images of Radboud dataset (target dataset). The main purpose behind utilizing a pre-trained model in this study is to counterbalance the relatively small size of the Radboud dataset. Consequently, this approach helps in mitigating the potential risk of overfitting. Therefore, we created a matrix with a size of 1,608 × 8,192 dimensions, where 1,608 signifies the quantity of images from the frontal Radboud dataset and 8,192 represents the scope of features, to encompass eight different emotions.

A multiclass SVM with a Gaussian kernel was used for the classification of all the experiments related to emotion recognition, and the parameters were optimised using the bayesopt optimiser [38] with ten-fold cross validation. Eight ROC curves and a confusion matrix for eight emotions were obtained, representing the performance of the classifier. Interestingly, this experiment demonstrated remarkable performance, reaching an accuracy of as high as 98.57%, which is the best achieved so far. The number of images was not that extensive in this experiment, and the results were

TABLE I. Description of all the public datasets of emotions utilized in this article.

Dataset	No. of images	Participants	Annotations	Condition
RadFD	8040	67 models	8 emotions	acted
KDEF	4900	70 individuals	7 emotions	acted
RAF-DB	29672	N/A	7 emotions	spontaneous
CelebA	30,000	10,177	40 attribute annotations	in the wild
CK+	327 seq. (10 to 60 frames/seq.)	210	8 emotions	posed + spontaneous

TABLE II. Area Under Curve (AUC) values for the eight experiments according to emotion recognition Experiments in section 4-A.

Emotions	Exp_1	Exp_2	Exp_3	Exp_4	Exp_5	Exp_6	Exp_7	Exp_8
Angry	0.982	0.999	0.959	0.951	**0.999**	0.937	0.888	0.759
Contemptuous	0.936	0.994	0.897	0.858	**0.999**	——	——	——
Disgusted	**1**	**1**	0.994	**1**	**1**	0.790	0.933	0.823
Fearful	**1**	0.998	0.977	0.972	0.998	0.925	0.723	0.835
Happy	**1**	**1**	0.998	0.993	**1**	0.833	0.974	0.719
Neutral	0.957	0.994	0.875	0.812	**0.999**	0.911	0.905	0.737
Sad	0.967	0.998	0.940	0.932	**0.999**	0.843	0.785	0.683
Surprised	**1**	0.999	0.991	**1**	**1**	0.851	0.943	0.720
Average	0.980	0.998	0.954	0.932	**0.999**	0.870	0.879	0.754

really promising. This comes as an advantage for leveraging the previous knowledge embedded inside the pre-trained model we use. See Exp1 in Table II, the ROC Curves of Exp1 for eight emotions in Fig 5, and the first confusion martix of Exp1 in Fig 6.

cutting-edge level of performance

2) In Exp2, a model pre-trained on the CelebA dataset was used to test the frontal Radboud dataset, to examine whether the model works better when it is trained on a large amount of data. The performance in Table II shows a significant improvement compared to Exp1. In experiment two, we utilized a previously trained model from the CelebA dataset to evaluate the performance of the frontal Radboud dataset. The aim was to investigate whether the model's effectiveness improves with extensive training on a vast quantity of data.

3) In Exp3, the performance of DCGAN was examined between frontal and multi-view images of emotions; a pre-trained model of the enhanced CK dataset was used to test the multi-view images of the Radboud. A matrix of feature vector of size 2,680 × 8,192 was produced. The recognition performance in Table II is comparatively lower when compared to only frontal Radboud images.

4) Exp4, another experiment, was conducted but here the trained model of the multi-view Radboud itself was used to test the multi-view Radboud images. It was found that the performance also significantly decreased and fell again.

5) Exp5 was conducted by training and testing on the frontal Radboud images using DCGAN. This experiment achieved promising results (an average AUC for all the emotions = 0.999, and accuracy = 97.64%), even with fewer images, for the same reasons mentioned above about the frontal Radboud dataset.

Finally, in 6), 7), 8) DCGAN was trained in the last three experiments (Exp6, Exp7, and Exp8) on three difficult datasets in the wild (RAF, KDEF, and SFEW), where facial expressions are close to the real-world environment. We can observe that the performance decreased significantly due to the apparent distortion of faces, low resolution imaging in the wild, and insufficient training data, specifically the SFEW image dataset, which limited the capacity to attain accurate results. This also could be attributed to other factors such as random background noise, clutter, head pose diversities, non-relevant variations and illumination changes, which are difficult to determine and might largely influence the DCGAN results. Furthermore, the categorization of emotions in natural environments is still a challenging issue that hampers performance. While DCGAN was not particularly developed for facial attribute extraction and classification, its outcomes in this context are encouraging.

The present study extensively employed the Receiver Operating Characteristic (ROC) curve to assess the best possible performance achieved by the specifically chosen classifier across different threshold settings in the tested models. This curve provides a graphical representation of the trade-off between true positive rates (sensitivity) and false positive rates (1-specificity), with sensitivity depicted on the y-axis and false alarm rate on the x-axis. A perfect classification scenario, wherein no misclassifications occur, is visually represented by a point in the top left corner of the plot. Conversely, a random classification yields a 45-degree diagonal line on the plot. The Area Under Curve (AUC) serves as a quantitative measure of the classifier's overall efficiency, with larger AUC values signifying superior performance. In this study, the ROC curves for each emotion in all eight experiments can be observed in Figure 8, while the accompanying AUC values are tabulated in Table 2.

The confusion matrices for eight facial expressions in all experiments were also calculated and shown in Fig 6. The correct classified unit for each expression is highlighted in dark blue, while the missclassified units were highlighted in paler blue. The experiments performed very well in recognising most of the emotions including: *surprise*,*fear*,*disgust*,*happiness*,*sadness*,*anger*,*contempt* and*neutral* with a true classification of 94.6% in Exp1 and Exp5. Also, sadness and disgust in Exp1, Exp2 had a correct classification of 100%. Anger and fear showed a relatively low recognition rate in experiments 1, 2, 3, and 4. Moreover, happiness and sadness expressions showed the lowest recognition rate of 38.5% and 38.8% in Exp8 respectively.Table III illustrates the accuracy achieved for each dataset in comparison to the highest performing techniques available. The values in the table were taken from the papers that introduced the methods and the experiments were different. There is a very legitimate and good point to be raised to explain the novelty of our approach and the advantage of this system compared to others, but it is mainly a limitation related to the adopted DCGAN itself, which we do not claim to propose in this work. The adopted DCGAN model, introduced by the authors in their original paper [6], is a powerful and versatile generative model. However, our work focuses on a different problem, specifically facial expression recognition in real-world conditions, using Deep Generative Adversarial Networks (DCGAN). This means that there may not be much transferable knowledge from their work to ours, as the challenges and objectives are distinct. Additionally, our research was hindered by the lack of a sufficiently large and diverse training dataset, as well as the difficulty of the datasets we chose, which closely mirror real-world facial expressions. These limitations restrained our ability to achieve high levels of accuracy in our results. To improve our method and offer a more equitable comparison with the most advanced methods available in Table III, we suggest two potential factors: the inclusion of large-scale datasets with comprehensive annotations that capture a wide range of facialdynamics,expressions,appearances,identities,and 3D pose variations, and the employment of a conditional DCGAN with ablation studies to assess the impact of these two factors on the accuracy of emotion recognition. This is important because emotion recognition holds great significance in the fields of computer vision and artificial intelligence, and it serves as a valuable benchmark for future research.

TABLE III. Analysis of the level of accuracy exhibited by each dataset in comparison to the current state-of-the-art approaches.

Dataset	Approach	Accuracy	Dataset	Approach	Accuracy
Radboud	(Ali et al.,2017 [39])	85.00%	SFEW	(Zhang et al.,2018 [21])	26.58%
	(Yaddaden et al.,2018 [40])	97.66%		(Dhall et al.,2015 [41])	35.93%
	(Jiang & Jia,2016 [42])	94.52%		(Levi & Hassner,2015 [43])	41.92%
	(Wu & Lin ,2018 [44])	96.27%		(Yao et al.,2015 [45])	44.04%
	(Mavani et al.,2017 [46])	95.71%		(Ng et al.,2015 [47])	48.50%
	(Sun et al.,2017 [48])	96.93%		(Yu & Zhang,2015 [49])	52.29%
	(C.Szegedy et al.,2015 [50])	95.45%		(Mollahosseini et al.,2016[51])	39.80&
	(Zavarez et al.,2017 [7])	85.97%		(Zhang et al.,2018[52])	55.27&
	(Li et al.,2019 [53])	96.11 %		(Mao et al.,2016 [54])	44.72%
	(WANG et al.,2019 [55])	80.69%		(Eleftheriadis et al.,2016 [56])	24.70 %
	ours	**98.57**%		**ours**	**44.52**%
KDEF	(Shin et al.,2016 [57])	59.15%	RAF	(Li et al.,2017 [33])	82.7%
	(Zavarez et al.,2017 [7])	72.55%		(Li et al.,2018 [58])	74.2%
	(Samara et al.,2019 [59])	81.84%		(Fan et al.,2018 [5])	76.73%
	(Yaddaden et al.,2018 [40])	79.69%		(Lin et al.,2018 [60])	75.73%
	(Ali et al.,2017 [39])	78.00%		(Ghosh et al.,2018 [61])	77.48%
	ours	**60.44**%		**ours**	**61.87**%

B. Experiments on Action Units (AUs)

In another set of experiments, we assessed how well the GAN features performed in recognizing individual AU.

1) Action Units on the Enhanced Cohn-Kanade Dataset

The objective of this experiment is to ascertain whether the features that have been acquired through learning, by the layer of a DCGAN and the Discriminator, can effectively capture and convey information that characterizes Action Units. To address this aim, the enhanced CK dataset, which offers comprehensive AU labeling, was employed. Fig 4, (a) and (b) indicate the original and generated images of the enhanced CK dataset. In this experiment (Exp.1), the 4D matrix was flattened and combined, resulting in dimensions of 8,422 × 8,192. These dimensions indicate that there were 8,422 images in the enhanced CK dataset with 8,129 feature vectors. We then trained and tested on the enhanced CK images using the linear SVM by the LibSVM [62] to identify the presence of 14 specific AUs (AU1,2,4,5,6,7,9,12,15,17,23,24,25,27); the findings from Exp.1, which can be found in section 4-B1, have been recorded in Table IV. In this table, you can see 14 different values for Areas Under the ROC Curve (AUC) corresponding to 14 different Action Units (AUs). The AUC values for the AUs from all the experiments in section 4-B1, 4-B2, and 4-B3 can also be found in Table IV. Additionally, Table V provides information on the pre-trained models that were utilized along with their respective datasets for evaluating the performance of cross dataset for AUs.

TABLE IV. AUC values for all the experiments regarding AUs shown in section 4-B.

AUs	Exp-1	Exp-2	Exp-3	Exp-4	Exp-5	Exp-6	Exp-7	Exp-8
AU1	**0.998**	0.994	0.961	0.909	0.896	0.996	0.890	0.896
AU2	0.918	**0.999**	0.948	0.744	0.633	0.998	0.705	0.677
AU4	0.788	**0.993**	0.887	0.515	0.656	0.990	0.519	0.663
AU5	0.982	**0.996**	0.991	0.676	0.808	0.980	0.624	0.767
AU6	0.895	**1**	0.908	0.574	0.518	**1**	0.568	0.542
AU7	**1**	0.984	**1**	0.757	0.649	0.979	0.642	0.601
AU9	0.990	**1**	0.997	0.668	0.546	**1**	0.683	0.565
AU12	**1**	0.990	0.980	0.515	0.633	0.967	0.549	0.615
AU15	0.932	**0.992**	0.963	0.510	0.659	0.990	0.518	0.647
AU17	0.785	**0.986**	0.873	0.708	0.807	0.984	0.623	0.748
AU23	0.882	**0.980**	0.927	0.865	0.828	0.979	0.775	0.814
AU24	0.948	**0.984**	0.902	0.849	0.688	0.979	0.822	0.721
AU25	0.998	**0.999**	0.992	0.799	0.929	0.999	0.755	0.894
AU27	0.672	**1**	0.669	0.668	0.541	0.999	0.678	0.519
Average	0.913	**0.993**	0.928	0.697	0.699	0.989	0.668	0.691

2) Radboud Emotions Relabelled to AUs

This experiment aims to evaluate whether features trained on a large (potentially unlabelled) dataset can be transferred for supervised training to a different one. This experiment was designed to confirm the results obtained on the CK dataset on a different dataset, namely by Radboud, since Radboud is only annotated for the eight basic emotions, and not for AU. The dataset was re-annotated according to the rules in [63].

While there have been numerous studies on AU detection, there is still limited research on effective approaches for associating AUs with emotions. The way to map emotions of the frontal Radboud dataset to AUs is summarised in Table VI. We utilized a pre-trained model from the enhanced CK dataset to extract the features of the frontal Radboud dataset. Following that, a linear SVM was employed for classification. The findings are highly intriguing; however, there were notable omissions in the crucial annotations concerning specific action units. These include AU10(upper lip raiser), AU11(nasolabial deepener), AU14(dimpler), AU20(lip stretcher), AU22(lip funneler), and AU26(jaw drop), which are commonly interpreted as an

TABLE V. Summaries all the pretrained models obtained from the DCGAN network with the related training and testing datasets regarding AUs.

Experiments	Pretrained models	Training	Testing
Exp.1	enhanced CK	enhanced CK	enhanced CK
Exp.2	enhanced CK	Radboud	Radboud
Exp.3	CelebA	enhanced CK	enhanced CK
Exp.4	CelebA	enhanced CK	Radboud
Exp.5	CelebA	Radboud	enhanced CK
Exp.6	CelebA	Radboud	Radboud
Exp.7	enhanced CK	enhanced CK	Radboud
Exp.8	enhanced CK	Radboud	enhanced CK

indication of happiness [64]. Furthermore, since contemptuous emotion (featuring AU 12 and 14 on one side of the face) is not recognized as one of Paul Ekman's six primary emotions, the Radboud dataset lacks specific guidelines for mapping it to action units. Instead, it represents a fusion of disgust and anger emotions. Also, there is no action unit to do lip corner tightening raised on only one side of a face. The results of Exp. 2, section 4-B2, in Tables IV and V, show the improvement in the results for all the AUs even with the imbalance and lowest occurrence activations in the dataset.

TABLE VI. A mapping between emotions and AUs based on rules according to the FACs [65].

Emotions	**AUs**
Happy	{AU6,AU12,AU25}
Sad	{AU1,AU4,AU17,AU15}
Fearful	{AU1,AU2,AU5,AU15,AU25}
Surprised	{AU1,AU2,AU5,AU25,AU27}
Angry	{AU4,AU5,AU7,AU17,AU23,AU24}
Disgusted	{AU9,AU15,AU17,AU25}
Contemptuous	{AU12}

3) *Transfer Learning on AUs*

The last experiment was conducted to evaluate the performance of cross-dataset evaluation research. More specifically, this involves using one dataset to train models and a different dataset to test them [58]. Transfer learning refers to the application of pre-trained models to address the inherent challenges stemming from the scarcity of data in a target dataset and to alleviate biases originating from uneven training sample sizes. A pre-trained model signifies a model that has been trained on an extensive benchmark dataset to tackle a different problem, albeit with a task that exhibits similarity and relevance to the specific problem being addressed. As the computational cost of training these models is substantial, it is customary in the field to adopt and employ models that have been rigorously documented and published in the literature. A pretrained model from the CelebA dataset, which means that the features in the DCGAN network were already learned (pre-training refers to the features in the DCGAN network), was utilized to train and test on the CK dataset. Subsequently, a pretrained model from the CelebA dataset was used to train and test on both the CK dataset and the Radboud dataset in a reciprocal manner. More information can be found in Experiments 3, 4, 5, 6, 7, and 8 (section 4-B3) in Table IV, Table V, and Figure 11. It is anticipated that the model achieved impressive outcomes when trained and evaluated on the same dataset, as demonstrated in Experiments 2 and 6 in Table IV. The performance of the cross-dataset was satisfactory, as shown in experiments 2, 3, and 6. For instance, the nose wrinkle (AU9) is commonly associated with disgust and occurs frequently, resulting in high areas under the curve (AUC) values of 1, 0.997, and 1 for these experiments respectively. Similarly, for lip parts (AU25), the AUC values are 0.999, 0.992, and 0.999. Additionally, the AUC value for lid lightener (AU7). In the cross-dataset performance of the CNN model, however, training and testing on two different datasets dropped the performance drastically because one of the datasets is quite different and fails to deal with new tasks and further operating settings that have not yet been seen during the training process and development. Notably, the results are encouraging for transferring *some* AUs. As we can observe from Table IV,AU1(inner brow raiser),AU23(lip tightener),AU24(lip pressor),and AU25 are transferred and generalized well for all experiments, while for the AU2 (outer brow raiser), and the AU17 (chin raiser) the performance is similar for all the values of AUCs in Exp.4,Exp.5,Exp.6,and Exp.8. The worst transfer appeared for AU4 (brow lowerer), with AUC = 0.515 in Exp.4 and AUC = 0.519 in Exp.7; AU4 is a common feature of confusion that happens on some occasions in our life, as well as AU6 (cheek raiser), AUC = 0.518 in Exp.5, AU12 (lip corner puller), AUC = 0.515 in Exp.4, AU15 (lip corner depressor), AUC = 0.510 in Exp.4, AUC = 0.518 in Exp.7, and AU27 (mouth stretch), AUC = 0.519 in Exp.8. The model exhibited optimal generalization performance in Experiments 2 and 6, achieving an average best prediction of 0.993 across all AUs. The second highest prediction accuracy observed was 0.989.

A smaller set of positive samples from the CK dataset, specifically AU7 (lid tightener), was used in an additional experiment to train the DCGAN. However, the performance of the model decreased, with an AUC of 0.58 for testing and 0.842 for training. This could be because AU7 is challenging to detect and distinguish from other AUs. Fig 8 in the paper displays some image samples of AU7 from the improved CK dataset.

Finally, one commonly employed qualitative method to assess the quality of generated samples in GANs is through human evaluation by visually examining the produced images. In our research, we have demonstrated that the DCGAN model offers significant improvements in training stability and effectively addresses the problem of mode collapse, all without introducing additional model complexity or compromising image quality. This improvement is discernible in Fig 4, which displays the generated facial expression images at varying resolutions using different datasets. In addition, the effectiveness of training the DCGAN model also relies on various factors such as

TABLE VII. A quantitave comparsion with the state of the art on RAFD dataset using IS, FID, SSIM and AMT metrics.

Model	IS (maximum is better) ↑	FID (lower is better) ↓	SSIM ↑	AMT ↑
Pix2Pix [66]	——	12.84	0.629	41.3%
pix2pixHD [67]	0.875	75.376	——	——
StarGAN [68]	1.036	56.937	0.8563	24.7%
GANimation [8]	1.112	34.360	0.8686	——
AF-VAE [69]	1.237	25.069	——	——
LEED [70]	——	38.20	0.8833	——
LGG + LS + TP [71]	——	**12.30**	0.705	74.9%
CycleGAN [72]	1.6942	52.8230	——	19.5%
Ours	**1.874**	22.318	**0.8942**	**76.72%**

the dataset's size, quality, quantity, and clarity. This study utilized the challenging RAF, KDEF, and SFEW datasets, all of which pose their difficulties. The chosen method has not suffered from such problems and can produce detectable images. Table VII presents a comprehensive analysis of the performance of generated image data samples using four prominent quantitative metrics: FID, IS, SSIM, and AMT. These metrics are considered state-of-the-art in evaluating the quality of samples on the RAFD dataset. The values reported in the table are sourced from diverse methodologies.

Mode collapse is a common issue that can occur when generating images using a neural network. It happens when the generator fails to accurately represent the variety and complexity of the training data. This leads to the generator producing limited and repetitive outputs, often resulting in low-quality images. Mode collapse is primarily attributed to the underlying difficulty encountered by the generator in adequately grasping and encapsulating the varied modes or patterns that are inherent in the training data. Consequently, the generator tends to produce outputs that manifest a notable degree of similarity or even complete identity, hence leading to a substantial deficiency in the diversity observed within the produced samples. To address this issue, there are several effective strategies. One approach that differs from using a single generator is to incorporate a collaborative or competitive framework among multiple generators. This allows each generator to specialize in capturing different modes of the data, resulting in a more comprehensive representation of the underlying distribution.

Another important factor is the selection of diverse training data. By including a wide range of samples that cover different modes and variations, we provide more opportunities for the generator to learn and reproduce the desired diversity in the generated outputs. Furthermore, choosing an appropriate architecture for the generator network is crucial. A well-designed network architecture can enhance the generator's ability to capture and express the complex and varied patterns present in the data. Overall, mitigating mode collapse requires a multi-faceted approach that combines diverse training data, multiple generators, and careful network architecture design. By considering and addressing these factors, we can enhance the overall quality and diversity of the produced images, ultimately promoting better generalization capabilities when dealing with unseen data. Figure 9, exemplifies an instance of mode collapse. The figure provided illustrates that although the model does not achieve perfect image generation, it exhibits the capacity to generate images which are perceptually discernible and recognizable by human observers. In recent times, human judgment has become widely adopted as a qualitative measure for evaluating GANs in terms of sample quality. Our research showcases the effectiveness of the DCGAN model in enhancing training stability and overcoming mode collapse without introducing additional complexity or compromising the quality of generated images. The substantial improvement becomes especially apparent when examining the high-resolution facial expression images generated from various datasets, as clearly illustrated in Figure 4.

Furthermore, the effectiveness and success of the DCGAN model rely on various crucial aspects such as the size, quality, quantity, and clarity of the training dataset samples utilized during the training process. In our study, we employed challenging datasets such as RAF, KDEF, and SFEW, all of which pose significant difficulties. However, our selected method successfully navigates these challenges and consistently generates identifiable images. We conducted a comprehensive assessment of the produced image data samples using four quantitative metrics: FID, IS, SSIM, and AMT. The evaluation results obtained from the RAFD dataset indicate a commendable level of performance that sets a new benchmark in the field.

5. Conclusions

The DCGAN network has been utilized as a highly efficient method for pre-training in the field of emotion recognition. The proposed method of identification has been experimentally validated on six standard datasets, effectively showcasing its advantageous performance across datasets of diverse sizes. This study concluded that training unsupervised DCGAN on a large-scale dataset produces powerful discriminative representation features for predicting and detecting AUs/emotions from frontal face images which is better than representing the multi-view of facial images. Additionally, it demonstrates that the suggested model possesses the capability to generalize. Future endeavours could encompass the training of a conditional DCGAN to separate the subject's facial expression from their identity. Furthermore, we intend to extend the current (2D) model to a (3D) counterpart by employing a conditional 3D-GAN, enabling the generation of videos.

References

[1] Y. Zhao, L. Yang, E. Pei, M. C. Oveneke, M. Alioscha-Perez, L. Li, D. Jiang, and H. Sahli, "Action unit driven facial expression synthesis from a single image with patch attentive gan," in *Computer Graphics Forum*. Wiley Online Library, 2021.

[2] H. Salehinejad, S. Valaee, T. Dowdell, E. Colak, and J. Barfett, "Generalization of deep neural networks for chest pathology classification in x-rays using generative adversarial networks," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 990–994.

[3] J. Han, Z. Zhang, N. Cummins, and B. Schuller, "Adversarial training in affective computing and sentiment analysis: Recent

advances and perspectives," *IEEE Computational Intelligence Magazine*, vol. 14, no. 2, pp. 68–81, 2019.

[4] D. Kollias, S. Cheng, E. Ververas, I. Kotsia, and S. Zafeiriou, "Deep neural network augmentation: Generating faces for affect analysis," *International Journal of Computer Vision*, vol. 128, no. 5, pp. 1455–1484, 2020.

[5] Y. Fan, J. C. Lam, and V. O. Li, "Multi-region ensemble convolutional neural network for facial expression recognition," in *International Conference on Artificial Neural Networks*. Springer, 2018, pp. 84–94.

[6] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[7] M. V. Zavarez, R. F. Berriel, and T. Oliveira-Santos, "Cross-database facial expression recognition based on fine-tuned deep convolutional network," in *2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. IEEE, 2017, pp. 405–412.

[8] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, "Ganimation: Anatomically-aware facial animation from a single image," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 818–833.

[9] X. Zhou, D. Dong, H. Wu, S. Zhao, D. Yu, H. Tian, X. Liu, and R. Yan, "Multi-view response selection for human-computer conversation," in *Proceedings of the 2016 conference on empirical methods in natural language processing*, 2016, pp. 372–381.

[10] J. Wang, "Improved facial expression recognition method based on gan," *Scientific Programming*, vol. 2021, pp. 1–8, 2021.

[11] J. F. Nash *et al.*, "Equilibrium points in n-person games," *Proceedings of the national academy of sciences*, vol. 36, no. 1, pp. 48–49, 1950.

[12] L. Alharbawee and N. Pugeault, "A benchmark of dynamic versus static methods for facial action unit detection," *The Journal of Engineering,*, vol. 2021, no. 5, pp. 252–266, 2021.

[13] M. R. Koujan, L. Alharbawee, G. Giannakakis, N. Pugeault, and A. Roussos, "Real-time facial expression recognition "in the wild" by disentangling 3d expression from identity," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 2020, pp. 24–31.

[14] A. Taherkhani, G. Cosma, and T. M. McGinnity, "Deep-fs: A feature selection algorithm for deep boltzmann machines," *Neurocomputing*, vol. 322, pp. 22–37, 2018.

[15] H. Larochelle and I. Murray, "The neural autoregressive distribution estimator," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011, pp. 29–37.

[16] X. Hou, L. Shen, K. Sun, and G. Qiu, "Deep feature consistent variational autoencoder," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2017, pp. 1133–1141.

[17] D. P. Kingma and M. Welling, "Stochastic gradient vb and the variational auto-encoder," in *Second International Conference on Learning Representations, ICLR*, vol. 19, 2014, p. 121.

[18] A. Spurr, E. Aksan, and O. Hilliges, "Guiding infogan with semi-supervision," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2017, pp. 119–134.

[19] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE transactions on affective computing*, 2020.

[20] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *International conference on machine learning*. PMLR, 2016, pp. 1558–1566.

[21] F. Zhang, T. Zhang, Q. Mao, and C. Xu, "Joint pose and expression modeling for facial expression recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3359–3368.

[22] J. Cao, Y. Hu, H. Zhang, R. He, and Z. Sun, "Learning a high fidelity pose invariant model for high-resolution face frontalization," *Advances in neural information processing systems*, vol. 31, 2018.

[23] J. T. Springenberg, "Unsupervised and semi-supervised learning with categorical generative adversarial networks," *arXiv preprint arXiv:1511.06390.*

[24] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3754–3762.

[25] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning gan for pose-invariant face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[26] X. Wang, X. Wang, and Y. Ni, "Unsupervised domain adaptation for facial expression recognition using generative adversarial networks," *Computational intelligence and neuroscience*, vol. 2018, 2018.

[27] R. Huang, S. Zhang, T. Li, and R. He, "Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2439–2448.

[28] H. Zhou, J. Sun, Y. Yacoob, and D. W. Jacobs, "Label denoising adversarial network (ldan) for inverse lighting of faces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6238–6247.

[29] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[30] Q. Ji, "Rpi intelligent systems lab (isl) image databases."

[31] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3730–3738.

[32] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. Van Knippenberg, "Presentation and validation of the radboud faces database," *Cognition and emotion*, vol. 24, no. 8, pp. 1377–1388, 2010.

[33] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild,"

in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2852–2861.

[34] D. Lundqvist, A. Flykt, and A. Öhman, "Karolinska directed emotional faces," *Cognition and Emotion*, 1998.

[35] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. IEEE, 2011, pp. 2106–2112.

[36] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[37] K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4193–42 032, 2017.

[38] R. Martinez-Cantin, "Bayesopt: a bayesian optimization library for nonlinear optimization, experimental design and bandits." *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3735–3739, 2014.

[39] A. M. Ali, H. Zhuang, and A. K. Ibrahim, "An approach for facial expression classification," *International Journal of Biometrics*, vol. 9, no. 2, pp. 96–112, 2017.

[40] Y. Yaddaden, M. Adda, A. Bouzouane, S. Gaboury, and B. Bouchard, "User action and facial expression recognition for error detection system in an ambient assisted environment," *Expert Systems with Applications*, vol. 112, pp. 173–189, 2018.

[41] A. Dhall, O. Ramana Murthy, R. Goecke, J. Joshi, and T. Gedeon, "Video and image based emotion recognition challenges in the wild: Emotiw 2015," in *Proceedings of the 2015 ACM on international conference on multimodal interaction*, 2015, pp. 423–426.

[42] B. Jiang and K. Jia, "Robust facial expression recognition algorithm based on local metric learning," *Journal of Electronic Imaging*, vol. 25, no. 1, p. 013022, 2016.

[43] G. Levi and T. Hassner, "Emotion recognition in the wild via convolutional neural networks and mapped binary patterns," in *Proceedings of the 2015 ACM on international conference on multimodal interaction*, 2015, pp. 503–510.

[44] B.-F. Wu and C.-H. Lin, "Adaptive feature mapping for customizing deep learning based facial expression recognition model," *IEEE access*, vol. 6, pp. 12 451–12 461, 2018.

[45] A. Yao, J. Shao, N. Ma, and Y. Chen, "Capturing au-aware facial features and their latent relations for emotion recognition in the wild," in *Proceedings of the 2015 acm on international conference on multimodal interaction*, 2015, pp. 451–458.

[46] V. Mavani, S. Raman, and K. P. Miyapuram, "Facial expression recognition using visual saliency and deep learning," in *Proceedings of the IEEE international conference on computer vision workshops*, 2017, pp. 2783–2788.

[47] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in *Proceedings of the 2015 ACM on international conference on multimodal interaction*, 2015, pp. 443–449.

[48] W. Sun, H. Zhao, and Z. Jin, "An efficient unconstrained facial expression recognition algorithm based on stack binarized autoencoders and binarized neural networks," *Neurocomputing*, vol. 267, pp. 385–395, 2017.

[49] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proceedings of the 2015 ACM on international conference on multimodal interaction*, 2015, pp. 435–442.

[50] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[51] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *2016 IEEE Winter conference on applications of computer vision (WACV)*. IEEE, 2016, pp. 1–10.

[52] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "From facial expression recognition to interpersonal relation prediction," *International Journal of Computer Vision*, vol. 126, no. 5, pp. 550–569, 2018.

[53] D. Li, Z. Li, R. Luo, J. Deng, and S. Sun, "Multi-pose facial expression recognition based on generative adversarial network," *IEEE Access*, vol. 7, pp. 143 980–143 989, 2019.

[54] Q. Mao, Q. Rao, Y. Yu, and M. Dong, "Hierarchical bayesian theme models for multipose facial expression recognition," *IEEE Transactions on Multimedia*, vol. 19, no. 4, pp. 861–873, 2016.

[55] X. Wang, Y. Wang, and W. Li, "U-net conditional gans for photo-realistic and identity-preserving facial expression synthesis," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 15, no. 3s, pp. 1–23, 2019.

[56] S. Eleftheriadis, O. Rudovic, and M. Pantic, "Discriminative shared gaussian processes for multiview and view-invariant facial expression recognition," *IEEE transactions on image processing*, vol. 24, no. 1, pp. 189–204, 2014.

[57] M. Shin, M. Kim, and D.-S. Kwon, "Baseline cnn structure analysis for facial expression recognition," in *2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN)*. IEEE, 2016, pp. 724–729.

[58] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 356–370, 2018.

[59] A. Samara, L. Galway, R. Bond, and H. Wang, "Affective state detection via facial expression analysis within a human–computer interaction context," *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, no. 6, pp. 2175–2184, 2019.

[60] F. Lin, R. Hong, W. Zhou, and H. Li, "Facial expression recognition with data augmentation and compact feature learning," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 1957–1961.

[61] S. Ghosh, A. Dhall, and N. Sebe, "Automatic group affect analysis in images via visual attribute and feature networks," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 1967–1971.

[62] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector

machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, pp. 1–27, 2011.

[63] J. Wu, Z. Huang, J. Thoma, D. Acharya, and L. Van Gool, "Wasserstein divergence for gans," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 653–668.

[64] M. E. Hoque, R. e. Kaliouby, and R. W. Picard, "When human coders (and machines) disagree on the meaning of facial affect in spontaneous videos," in *International Workshop on Intelligent Virtual Agents*. Springer, 2009, pp. 337–343.

[65] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proceedings of the National Academy of Sciences*, vol. 111, no. 15, pp. E1454–E1462, 2014.

[66] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[67] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8798–8807.

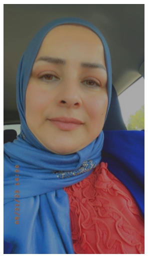
[68] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8789–8797.

[69] S. Qian, K.-Y. Lin, W. Wu, Y. Liu, Q. Wang, F. Shen, C. Qian, and R. He, "Make a face: Towards arbitrary high fidelity face manipulation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10 033–10 042.

[70] R. Wu and S. Lu, "Leed: Label-free expression editing via disentanglement," in *European Conference on Computer Vision*. Springer, 2020, pp. 781–798.

[71] J. Zhang, X. Zeng, Y. Pan, Y. Liu, Y. Ding, and C. Fan, "Faceswapnet: Landmark guided many-to-many face reenactment," *arXiv preprint arXiv:1905.11805*, vol. 2, p. 3, 2019.

[72] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.



Luma Alharbawee I am a faculty member in the College of Computer Sciences and Mathematics, Department of Statistics and Informatics, University of Mosul, Mosul, Iraq. I have a master's degree in Computer Science in the field of Digital Image Processing, from the University of Mosul in 2002. I obtained my Ph.D. in Artificial intelligence in 2019, from the University of Exeter, College of Engineering, Mathematics and Physical Sciences- Department of Computer Science. Exeter, Devon, England, United Kingdom. My current Research interests are in the field of Cognitive Vision and Machine Learning.

Nicolas Pugeault After finishing my Engineer studies at the ESIEA Paris, I moved to the UK and did an MSc in Computational Intelligence at the University of Plymouth, before moving to Stirling to study with Norbert Krueger and Florentin Woergoetter on the European project ECOVISION. I graduated my PhD from the University of Goettingen in 2008. From September 2007 to September 2009, I held a half-time Assistant Professor position at the University of Southern Denmark, and a half-time Research Associate position at the University of Edinburgh, collaborating with Prof. Norbert Krueger on the two EU projects DrivSco and PACOPLUS. I moved to the Centre for Vision Speech and Signal Processing (CVSSP) at the University of Surrey in 2009 as a Postdoc, working with Prof. Richard Bowden on autonomous driving and automatic sign language recognition (I was involved with the projects DIPLECS and Dicta-Sign), and was offered a lectureship in 2013. In 2016, I moved to the Department of Computer Science at the University of Exeter, where my group's research focused on the use of deep neural networks for modelling visual attention and context in visual scenes (EPSRC grant DEVA). I started as a Reader in the School of Computing at the University of Glasgow in June 2020.

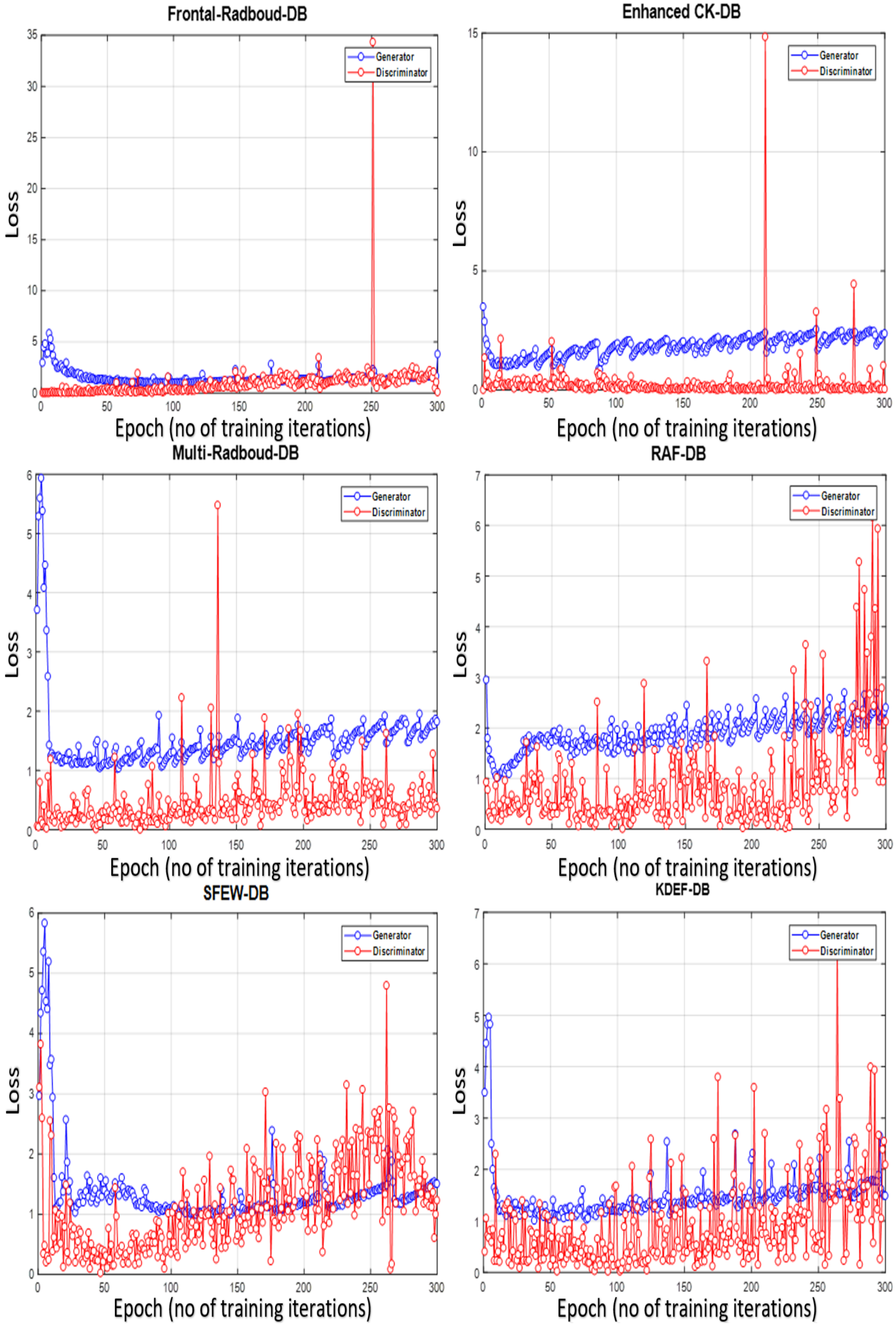


Figure 3. The logistic loss and convergence of the G and D during training DCGAN under multiple iterations on the datasets used in this work: (a) frontal Radboud, (b) Enhanced CK, (c) multi Radboud, (d) RAF, (e) SFEW, (f) KDEF dataset.

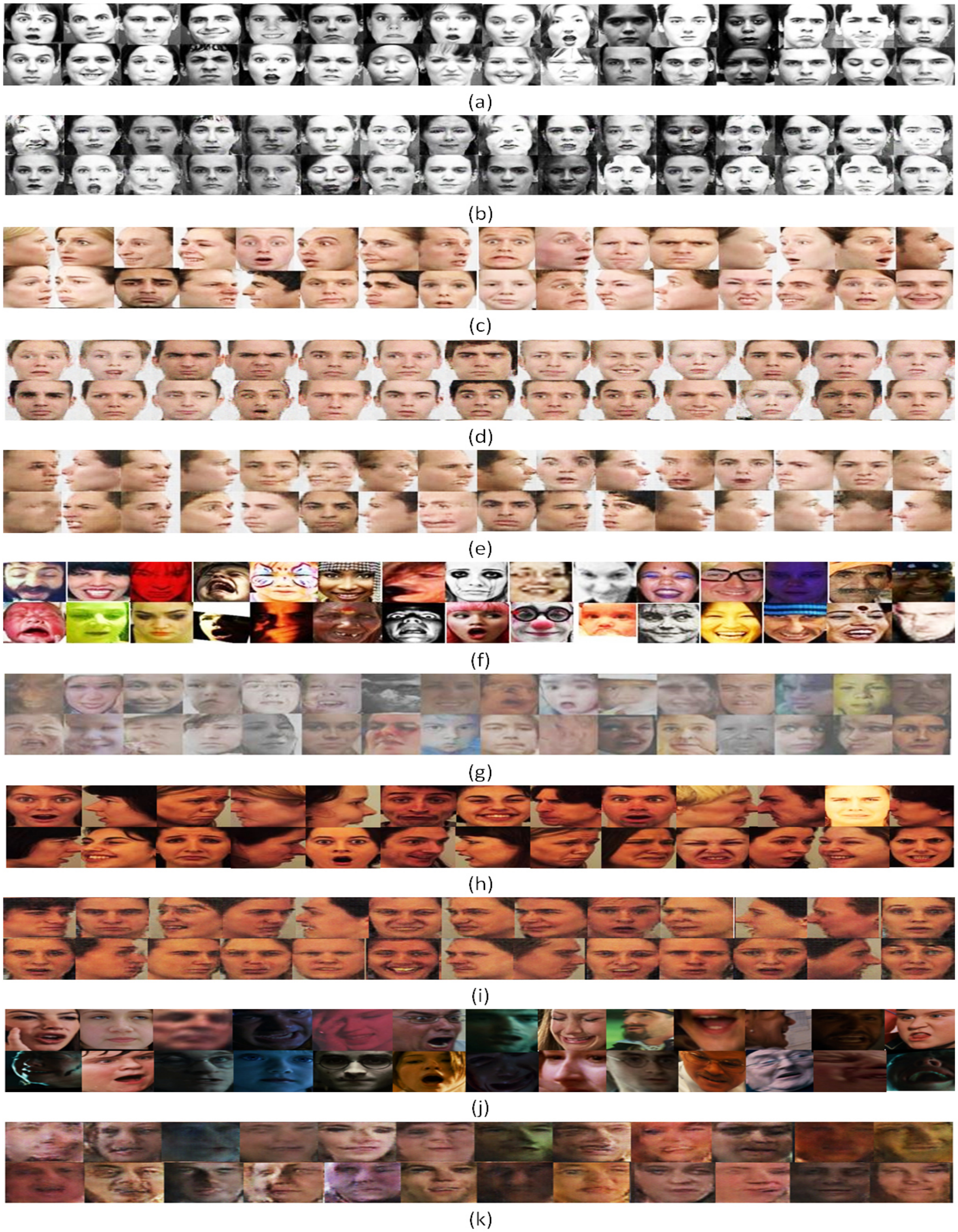


Figure 4. Comparison of the selected samples of the generated facial expression images using DCGAN on different datasets: (a) & (b) enhanced Cohn-Kanade original and generated images, (c) original frontal and multi-view Radboud images, (d) frontal Radboud generated images, (e) multi-view Radboud generated images, (f) & (g) RAF original and generated images, (h) & (i) KDEF original and generated images, (j) & (k) SFEW original and generated images.

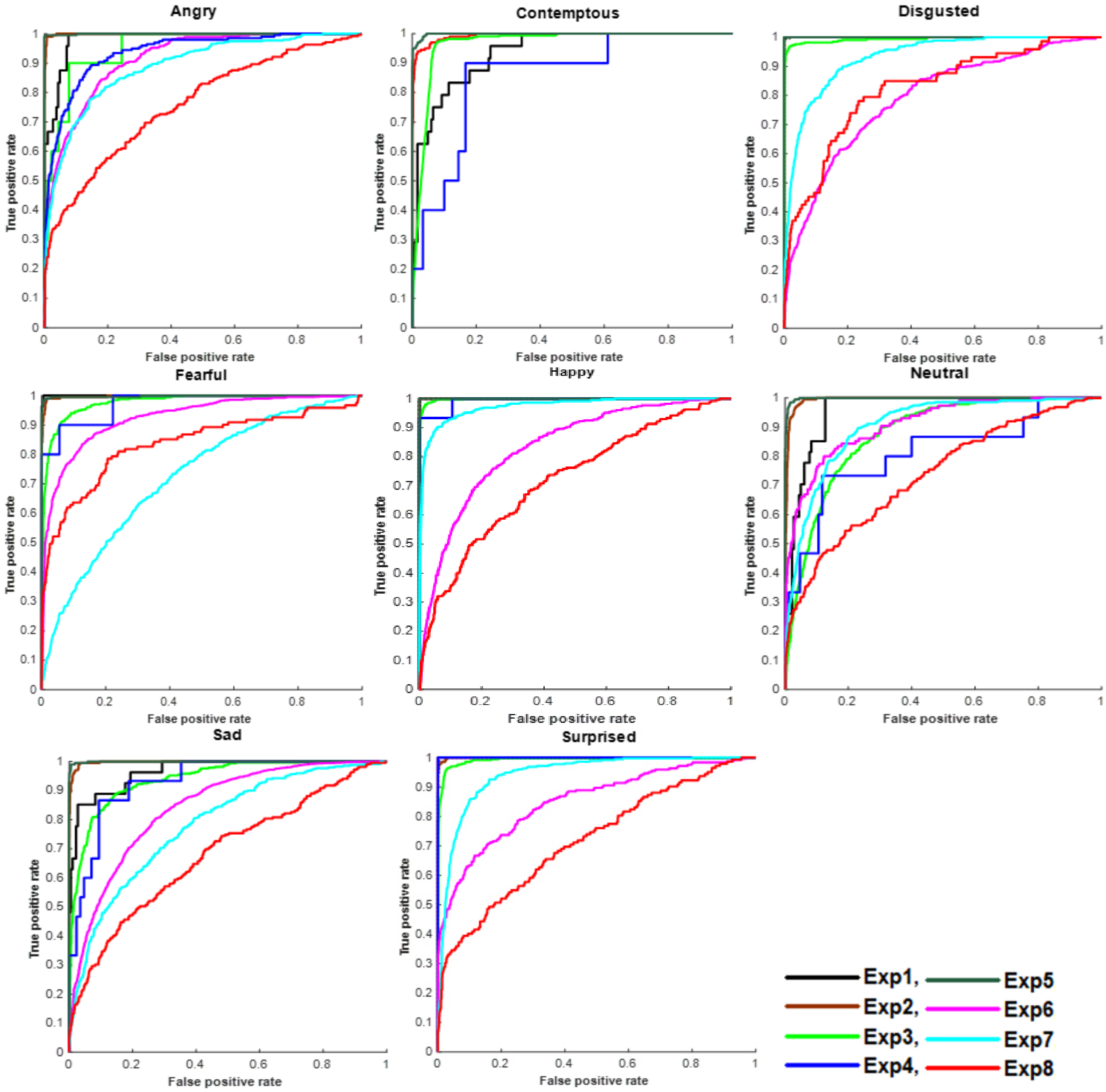
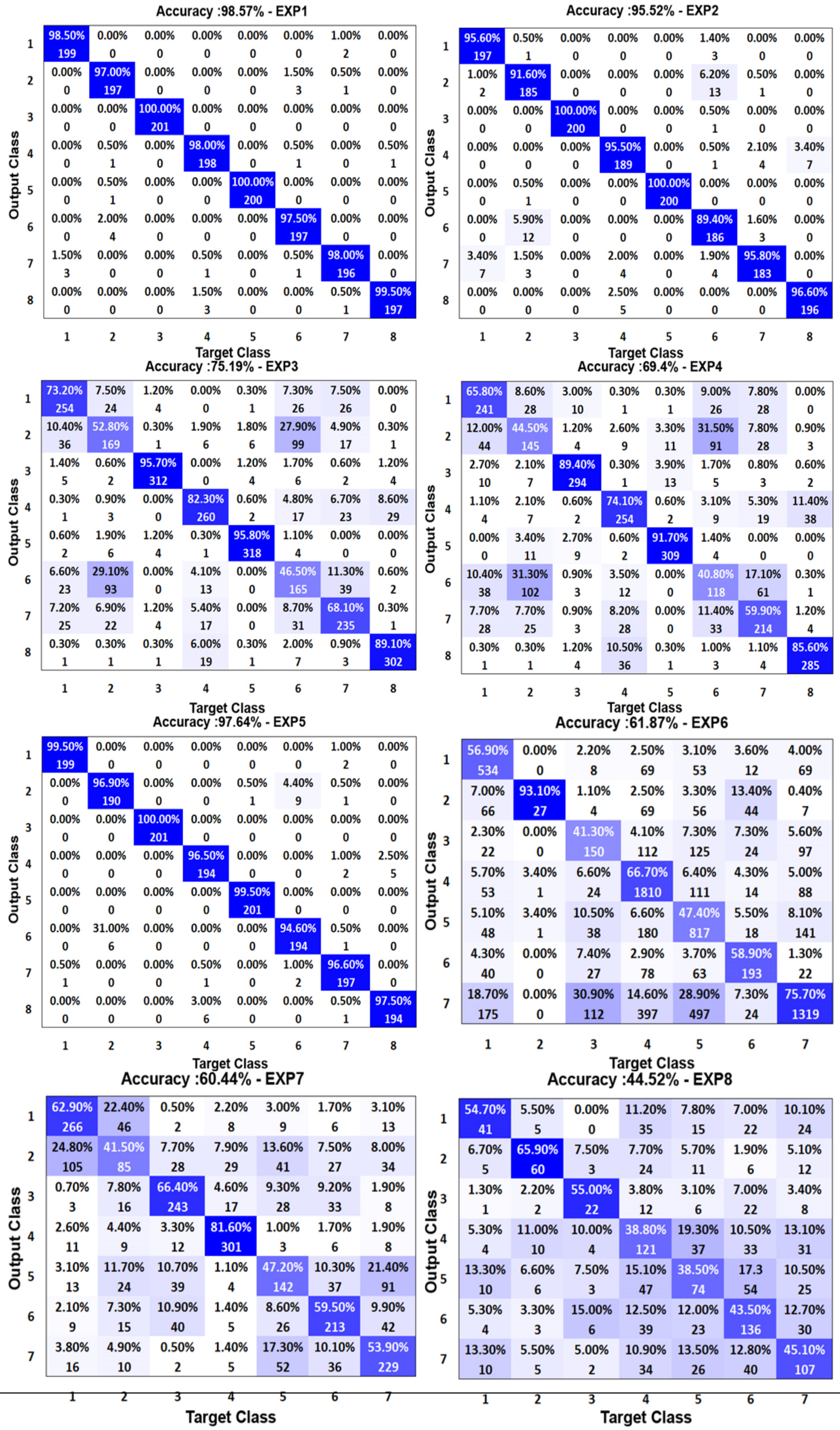


Figure 5. Receiver Operating Curves (ROC) for eight emotions and eight experiments; each figure depicts ROC curve of eight dissimilar experiments.



Accuracy :98.57% - EXP1

Output Class / Target Class

	1	2	3	4	5	6	7	8
1	98.50% 199	0.00% 0	0.00% 0	0.00% 0	0.00% 0	0.00% 0	1.00% 2	0.00% 0
2	0.00% 0	97.00% 197	0.00% 0	0.00% 0	0.00% 0	1.50% 3	0.50% 1	0.00% 0
3	0.00% 0	0.00% 0	100.00% 201	0.00% 0	0.00% 0	0.00% 0	0.00% 0	0.00% 0
4	0.00% 0	0.50% 1	0.00% 0	98.00% 198	0.00% 0	0.50% 1	0.00% 0	0.50% 1
5	0.00% 0	0.50% 1	0.00% 0	0.00% 0	100.00% 200	0.00% 0	0.00% 0	0.00% 0
6	0.00% 0	2.00% 4	0.00% 0	0.00% 0	0.00% 0	97.50% 197	0.00% 0	0.00% 0
7	1.50% 3	0.00% 0	0.00% 0	0.50% 1	0.00% 0	0.50% 1	98.00% 196	0.00% 0
8	0.00% 0	0.00% 0	0.00% 0	1.50% 3	0.00% 0	0.00% 0	0.50% 1	99.50% 197

Accuracy :95.52% - EXP2

Output Class / Target Class

	1	2	3	4	5	6	7	8
1	95.60% 197	0.50% 1	0.00% 0	0.00% 0	0.00% 0	1.40% 3	0.00% 0	0.00% 0
2	1.00% 2	91.60% 185	0.00% 0	0.00% 0	0.00% 0	6.20% 13	0.50% 1	0.00% 0
3	0.00% 0	0.00% 0	100.00% 200	0.00% 0	0.00% 0	0.50% 1	0.00% 0	0.00% 0
4	0.00% 0	0.00% 0	0.00% 0	95.50% 189	0.00% 0	0.50% 1	2.10% 4	3.40% 7
5	0.00% 0	0.50% 1	0.00% 0	0.00% 0	100.00% 200	0.00% 0	0.00% 0	0.00% 0
6	0.00% 0	5.90% 12	0.00% 0	0.00% 0	0.00% 0	89.40% 186	1.60% 3	0.00% 0
7	3.40% 7	1.50% 3	0.00% 0	2.00% 4	0.00% 0	1.90% 4	95.80% 183	0.00% 0
8	0.00% 0	0.00% 0	0.00% 0	2.50% 5	0.00% 0	0.00% 0	0.00% 0	96.60% 196

Accuracy :75.19% - EXP3

Output Class / Target Class

	1	2	3	4	5	6	7	8
1	73.20% 254	7.50% 24	1.20% 4	0.00% 0	0.30% 1	7.30% 26	7.50% 26	0.00% 0
2	10.40% 36	52.80% 169	0.30% 1	1.90% 6	1.80% 6	27.90% 99	4.90% 17	0.30% 1
3	1.40% 5	0.60% 2	95.70% 312	0.00% 0	1.20% 4	1.70% 6	0.60% 2	1.20% 4
4	0.30% 1	0.90% 3	0.00% 0	82.30% 260	0.60% 2	4.80% 17	6.70% 23	8.60% 29
5	0.60% 2	1.90% 6	1.20% 4	0.30% 1	95.80% 318	1.10% 4	0.00% 0	0.00% 0
6	6.60% 23	29.10% 93	0.00% 0	4.10% 13	0.00% 0	46.50% 165	11.30% 39	0.60% 2
7	7.20% 25	6.90% 22	1.20% 4	5.40% 17	0.00% 0	8.70% 31	68.10% 235	0.30% 1
8	0.30% 1	0.30% 1	0.30% 1	6.00% 19	0.30% 1	2.00% 7	0.90% 3	89.10% 302

Accuracy :69.4% - EXP4

Output Class / Target Class

	1	2	3	4	5	6	7	8
1	65.80% 241	8.60% 28	3.00% 10	0.30% 1	0.30% 1	9.00% 26	7.80% 28	0.00% 0
2	12.00% 44	44.50% 145	1.20% 4	2.60% 9	3.30% 11	31.50% 91	7.80% 28	0.90% 3
3	2.70% 10	2.10% 7	89.40% 294	0.30% 1	3.90% 13	1.70% 5	0.80% 3	0.60% 2
4	1.10% 4	2.10% 7	0.60% 2	74.10% 254	0.60% 2	3.10% 9	5.30% 19	11.40% 38
5	0.00% 0	3.40% 11	2.70% 9	0.60% 2	91.70% 309	1.40% 4	0.00% 0	0.00% 0
6	10.40% 38	31.30% 102	0.90% 3	3.50% 12	0.00% 0	40.80% 118	17.10% 61	0.30% 1
7	7.70% 28	7.70% 25	0.90% 3	8.20% 28	0.00% 0	11.40% 33	59.90% 214	1.20% 4
8	0.30% 1	0.30% 1	1.20% 4	10.50% 36	0.30% 1	1.00% 3	1.10% 4	85.60% 285

Accuracy :97.64% - EXP5

Output Class / Target Class

	1	2	3	4	5	6	7	8
1	99.50% 199	0.00% 0	0.00% 0	0.00% 0	0.00% 0	0.00% 0	1.00% 2	0.00% 0
2	0.00% 0	96.90% 190	0.00% 0	0.00% 0	0.50% 1	4.40% 9	0.50% 1	0.00% 0
3	0.00% 0	0.00% 0	100.00% 201	0.00% 0	0.00% 0	0.00% 0	0.00% 0	0.00% 0
4	0.00% 0	0.00% 0	0.00% 0	96.50% 194	0.00% 0	0.00% 0	1.00% 2	2.50% 5
5	0.00% 0	0.00% 0	0.00% 0	0.00% 0	99.50% 201	0.00% 0	0.00% 0	0.00% 0
6	0.00% 0	31.00% 6	0.00% 0	0.00% 0	0.00% 0	94.60% 194	0.50% 1	0.00% 0
7	0.50% 1	0.00% 0	0.00% 0	0.50% 1	0.00% 0	1.00% 2	96.60% 197	0.00% 0
8	0.00% 0	0.00% 0	0.00% 0	3.00% 6	0.00% 0	0.00% 0	0.50% 1	97.50% 194

Accuracy :61.87% - EXP6

Output Class / Target Class

	1	2	3	4	5	6	7
1	56.90% 534	0.00% 0	2.20% 8	2.50% 69	3.10% 53	3.60% 12	4.00% 69
2	7.00% 66	93.10% 27	1.10% 4	2.50% 69	3.30% 56	13.40% 44	0.40% 7
3	2.30% 22	0.00% 0	41.30% 150	4.10% 112	7.30% 125	7.30% 24	5.60% 97
4	5.70% 53	3.40% 1	6.60% 24	66.70% 1810	6.40% 111	4.30% 14	5.00% 88
5	5.10% 48	3.40% 1	10.50% 38	6.60% 180	47.40% 817	5.50% 18	8.10% 141
6	4.30% 40	0.00% 0	7.40% 27	2.90% 78	3.70% 63	58.90% 193	1.30% 22
7	18.70% 175	0.00% 0	30.90% 112	14.60% 397	28.90% 497	7.30% 24	75.70% 1319

Accuracy :60.44% - EXP7

Output Class / Target Class

	1	2	3	4	5	6	7
1	62.90% 266	22.40% 46	0.50% 2	2.20% 8	3.00% 9	1.70% 6	3.10% 13
2	24.80% 105	41.50% 85	7.70% 28	7.90% 29	13.60% 41	7.50% 27	8.00% 34
3	0.70% 3	7.80% 16	66.40% 243	4.60% 17	9.30% 28	9.20% 33	1.90% 8
4	2.60% 11	4.40% 9	3.30% 12	81.60% 301	1.00% 3	1.70% 6	1.90% 8
5	3.10% 13	11.70% 24	10.70% 39	1.10% 4	47.20% 142	10.30% 37	21.40% 91
6	2.10% 9	7.30% 15	10.90% 40	1.40% 5	8.60% 26	59.50% 213	9.90% 42
7	3.80% 16	4.90% 10	0.50% 2	1.40% 5	17.30% 52	10.10% 36	53.90% 229

Accuracy :44.52% - EXP8

Output Class / Target Class

	1	2	3	4	5	6	7
1	54.70% 41	5.50% 5	0.00% 0	11.20% 35	7.80% 15	7.00% 22	10.10% 24
2	6.70% 5	65.90% 60	7.50% 3	7.70% 24	5.70% 11	1.90% 6	5.10% 12
3	1.30% 1	2.20% 2	55.00% 22	3.80% 12	3.10% 6	7.00% 22	3.40% 8
4	5.30% 4	11.00% 10	10.00% 4	38.80% 121	19.30% 37	10.50% 33	13.10% 31
5	13.30% 10	6.60% 6	7.50% 3	15.10% 47	38.50% 74	17.3 54	10.50% 25
6	5.30% 4	3.30% 3	15.00% 6	12.50% 39	12.00% 23	43.50% 136	12.70% 30
7	13.30% 10	5.50% 5	5.00% 2	10.90% 34	13.50% 26	12.80% 40	45.10% 107

Figure 6. The average recognition rate of a Confusion matrix is obtained from SVM classifier for the eight experiments.

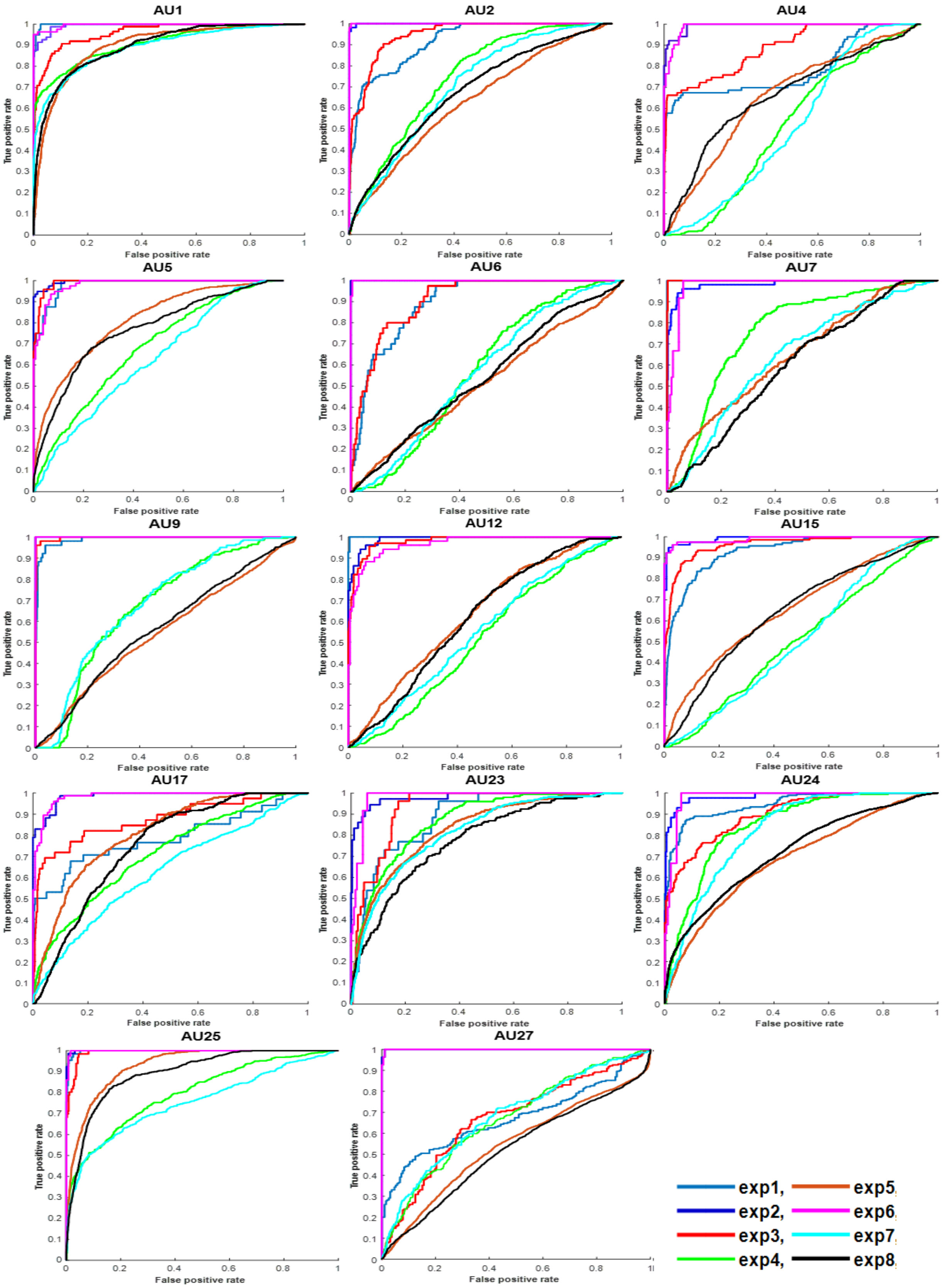


Figure 7. Receiver Operating Curves (ROC) for fourteen Action Units (AU) and eight experiments, each figure depicts ROC curve of eight dissimilar experiments.

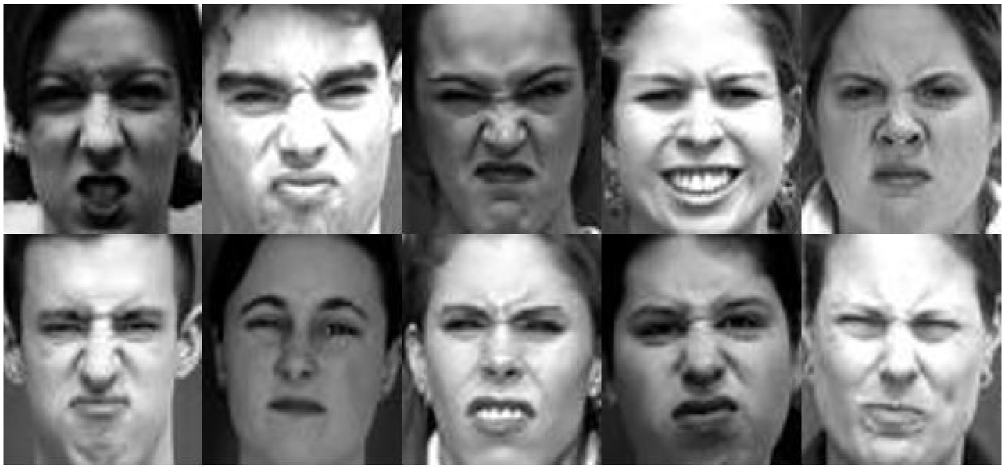


Figure 8. Images from the enhanced CK dataset represent AU7.

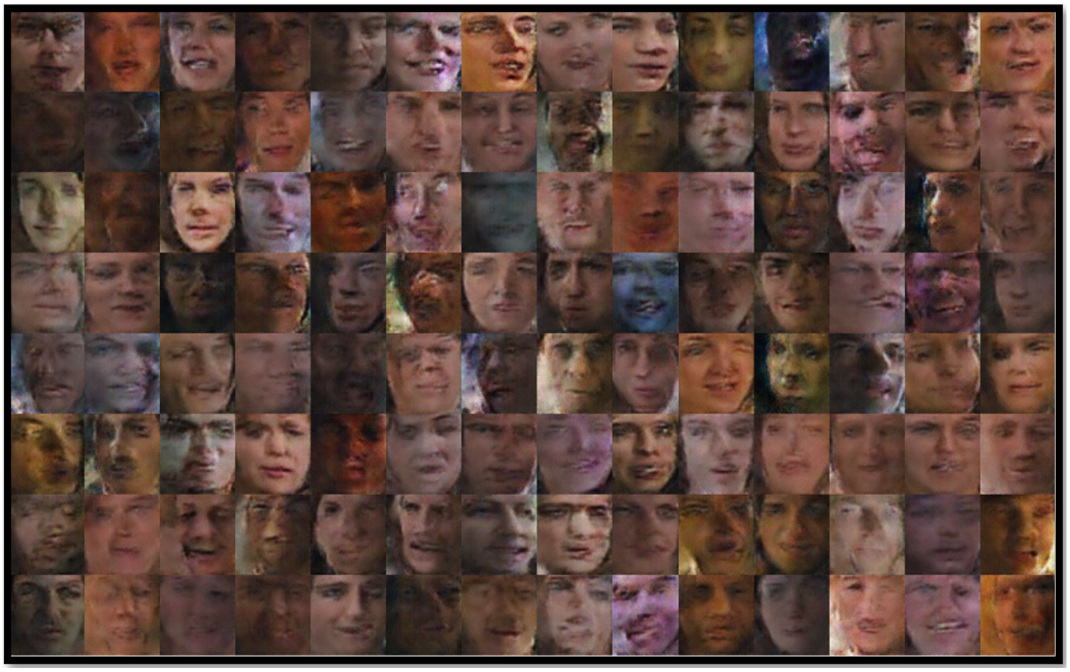


Figure 9. An example of mode collapse of the generated images.