IJCNN 2019. International Joint Conference on Neural Networks. Budapest, Hungary. 14-19 July 2019

Multiple Fake Classes GAN for Data Augmentation in Face Image Dataset

Adamu Ali-Gombe[1] Eyad Elyan[1] Chrisina Jayne[2]
[1]Robert Gordon University [2]Oxford Brookes University

***Abstract*— Class-imbalanced datasets often contain one or more class that are under-represented in a dataset. In such a situation, learning algorithms are often biased toward the majority class instances. Therefore, some modification to the learning algorithm or the data itself is required before attempting a classification task. Data augmentation is one common approach used to improve the presence of the minority class instances and rebalance the dataset. However, simple augmentation techniques such as applying some affine transformation to the data, may not be sufficient in extreme cases, and often do not capture the variance present in the dataset. In this paper, we propose a new approach to generate more samples from minority class instances based on Generative Adversarial Neural Networks (GAN). We introduce a new Multiple Fake Class Generative Adversarial Networks (MFC-GAN) and generate additional samples to rebalance the dataset. We show that by introducing multiple fake class and oversampling, the model can generate the required minority samples. We evaluate our model on face generation task from attributes using a reduced number of samples in the minority class. Results obtained showed that MFC-GAN produces plausible minority samples that improve the classification performance compared with state-of-the-art AC-GAN generated samples.**

I. Introduction

Imbalanced-class datasets need to be handled prior to fitting a learning algorithm in order to avoid biased models. This is a common problem across many domains including the vision domain. For example, 2D image generation and classification such as face verification [14], face reading [22] and activity recognition using facial images [26]. Such application requires large volumes of training data with adequate number of samples per class. Therefore, an imbalanced face-images dataset needs to be handled prior to training.

Methods for handling class-imbalance can be broadly categorised as data-level solutions or algorithm-level solutions. Under-sampling and oversampling are re-sampling commonly used to handle class imbalanced datasets. These methods are simple and often yield better results. However, the main drawback of applying such sampling methods is the potential loss of data which may be caused by under-sampling or possible overfitting due to oversampling [4]. Algorithm-based solutions on the other hand aims at modifying the learning methods to account for the skewed class distribution. This includes adjusting the cost function to account to for misclassifying the class of interest [13].

Another common data-level approach is based on data augmentation which aims at synthesizing more samples from existing dataset [6]. Such methods proved to reduce the effect of class imbalance and improve generalisation of the model [3]. However, in extreme imbalance cases augmentation may fail to produce enough variations in samples. Moreover, augmenting face image attributes like hair colour, gender, eyeglasses or smile; presents a more challenging task for simple augmentation techniques [20]. A realistic approach will be to train a generative model that can capture these facial attributes while generating plausible samples that are suitable for augmentation.

The recent advances in generative modelling have seen the generation of high-resolution images with high fidelity that are indifferent from the original training data. This development created an opportunity to resolve the problem of bias and class imbalance problem in datasets by generating synthetic samples for augmentation. The challenge here is that a generative model trained on a class imbalance dataset may not necessarily capture the actual data distribution, especially in extreme conditions where enough samples are not available to train the models. Generative Adversarial Networks (GANs) are state-of-the-art in image generation and Conditional GAN (C-GAN) [17] offers a class-specific sample generation from labels. However, GAN like other neural networks when trained on imbalance classes is affected by this problem. For instance, both Auxiliary Classifier GAN (AC-GAN) [19], C-GAN and GAN [8] avoid generating minority classes in extreme class imbalance cases [16].

In this paper, we propose a novel Multiple Fake Classes GAN model (MFC-GAN) for performing image generation from underrepresented class instances. Few Shot Classifier GAN (FSC-GAN) [1] implemented multiple fake classes but the samples generated were not suitable for augmentation. This is because FSC-GAN sample had artefacts or white noise patches in images. Moreover, a preliminary investigation revealed that FSC-GAN suffers from class imbalance problem and fails to generate minority class samples in extreme cases. MFC-GAN generates class-specific samples through generator conditioning. In addition, the model implements a classifier which isolates real samples into real classes and fake samples into multiple fake classes. MFC-GAN is trained with a modified objective, and we demonstrate that class imbalance problem could be addressed through re-sampling of the minority class. The proposed model was evaluated on face image generation problem from face image attributes that are under-represented in the CelebA dataset [15] which is a widely used dataset for face generation and classification [10]. CelebA attributes are represented as binary labels describing the presence or absence of a face feature such as a beard or

978-1-7281-2009-6/$31.00 paper N-20152.pdf

no beard. The images distribution across these attributes varies significantly thereby creating a class imbalance problem. We explored this problem and created more scenarios by reducing the number of samples in the minority classes. Our experiments considered two minority classes namely, eyeglasses and goatee attributes. We study the generation of samples from reduced number of instances. Furthermore, the generated samples were used as augmentation set to rebalance the classes and improve classification performance. We compare MFC-GAN performance in generating additional minority samples to state-of-the-art AC-GAN. The results obtained showed that MFC-GAN performs better than AC-GAN model on different classification metrics.

Our contributions in this paper are as follows:

- Multiple Fake Classes GAN model (MFC-GAN) for generating additional samples of face images with specific facial attributes in extreme imbalance scenarios
- Apply the generated samples to rebalance the dataset and improve classification of the under-represented attributes
- Our method shows that it is possible to train Deep GAN models using a small number of examples. This opens new promising research directions on training Deep Models with smaller sets of data

The remainder of this paper is organized as follows. In Section II, literature is reviewed. Section III presents the proposed method. Section IV discusses in details experimental set-up, the dataset, and the results are presented in section V. Findings are discussed in section VI. Finally, we draw conclusions and suggest future directions in Section VII.

II. Related Works

Facial attribute classification is challenging because they vary significantly from one person to another [5]. Face pose angles, different lighting conditions and variety of clothing such as eyeglasses, caps, and jewelry can create an occlusion. Furthermore, an imbalance in facial attribute classes makes the classification task challenging. Attribute classification approaches can be grouped into two categories. The first category considers the local image patches by feeding in outputs from attribute detectors. The problem with this group is the sole reliance on the efficiency of the detection model [5]. And the second approaches process the global image to extract the required features and classify attributes. The latter methods are more robust and have provided state-of-the-art performances recently. Furthermore, global approaches have been implemented as multi-tasking approaches [5] and in some cases employing specific models to classify each attribute [15], [25]. More recently, multi-task models have explored the correlation between attributes to improve classification performances as shown in [11].

Multi-task approaches are mostly multi-model and utilize shared information from related problems. For instance, [15] proposed the use of a multi-model framework to perform attributes classification. The framework consists of a face localization network (LNet) and an attribute classification network (ANet) that feeds on a localized face from LNet. Pre-training both networks differently on a face recognition task proved to be more efficient than training from scratch. In this framework, a different SVM is trained for each attribute using the features extracted. A multi-model approach implements a different model to classify each attribute which may be cumbersome when attribute classes are large. However, it can be effective when considering specific characteristics like smile and gender. Zhang *et al.* [25] learn to classify gender and smile attributes from facial images using two separate networks (GNet and SNet). An exciting part of the study is the use of the correlation between specific attributes to improve performance in low data regimes. The two models were pre-trained on VGG-Faces and CelebA dataset before fine tuning on FotW dataset in a general-to-specific manner.

Generating specific facial feature in images has many desirable applications such as security, fashion and in supporting other processes such as classification. Conditional GANs (C-GAN) provides the required functionality to generate faces with specific attributes. For instance, Jon in [7] used a variant of conditional GANs to generate faces from attribute vectors. Attributes were used to condition the generator and discriminator, but the author was able to control limited attributes combinations. AC-GANs, on the other hand, possesses some characteristics of C-GAN specifically conditional image generation. An extra classification task in AC-GAN re-enforces class-specific generation and improve sample quality and diversity. Research into this area revealed that face generation with auxiliary classification frameworks mostly rely on a hybrid approach using an auto-encoder model to learn or extract features before the GAN model is trained. For instance, Fine-grained Multi-attribute GAN (FM-GAN) [24] was used to generate plausible faces with precise age using facial attributes. The model is a modified AC-GAN that incorporates attributes into the generator. The authors used the conditional reconstruction of the embeddings and considered three sets of attributes: age, gender, and ethnicity. FM-GAN was trained on CelebA dataset and the synthesized images that were used to augment MORPH II dataset[1]. The new dataset was evaluated using a Convolutional Neural Network (CNN) and results obtained showed that the classifier performs better when the synthetic samples were added.

A similar strategy was used also used by Balancing GAN (BAGAN) [16]. Both the decoder and encoder were learned before adversarial training and were used to initialize BAGAN generator and discriminator respectively. This prevented the model from mode collapse and generated quality samples that were used in augmenting an imbalanced dataset. At GAN training phase, the latent vector is class-conditioned and sampled from the learned z-space from auto-encoder training. A final soft-max layer is also added to the discriminator with an extra class for all generated samples (n+1). Results show better classification results were obtained by augmenting with BAGAN and samples had better Structural Similarity Index

[1]https://ebill.uncw.edu/C20231_ustores/web/classic/store_main.jsp?STOREID=4

(SSIM) [19] than both AC-GAN and the original GAN.

In the medical domain, Deep Convolution GAN (DCGAN) was used in [6] to synthesize liver lesion samples which were used to enhance CNN performance through data augmentation. The original dataset is small and training the GAN model with few examples was possible through the application of traditional augmentation techniques. Closely related to this is MelanoGAN [3], which is a DCGAN coupled with a Laplacian GAN (LAPGAN). MelanoGAN was used to synthesize more skin lesion samples to reduce the effect of data imbalance in training ResNet-50 classifier. In face recognition domain, Data Augmentation GAN (DAGAN) [2] generated plausible facial images that were used in training matching networks [23]. The generated samples were used as augmentation samples to improve the performance of the face recognition model. To achieve this, DAGAN model requires to be trained separately before a sample selector network is collectively trained with the model on the final training phase.

All these models in the literature share some similarity with our approach, that is synthesizing more samples for augmentation. However, we use a different GAN model with multiple fake classes, and we investigate face generation from attributes that are under-represented. CelebA dataset is one of the most widely used benchmarks for facial attributes classification and face generation. While significant achievements have been recorded on this dataset, some interesting potentials still remain untapped. Hand *et al.* [10] pointed out that the dataset is biased towards posed celebrity images that are not indicative of the real world. Looking at the attribute distribution across images, we can see that the dataset is biased toward frontal faces, smiling and mostly young celebrity pictures. The authors argued that models trained on this dataset without putting into account such biases might perform poorly on a different domain. And balancing by re-sampling a class directly affect other class distribution as well [10].

III. Method

The proposed approach trains a multiple fake classes GAN on a class imbalance dataset of face images. The trained model is used to generate plausible samples from the minority class examples. The generated minority samples are then used to augment the original training set to rebalance the dataset. Finally, we validate the approach on classification task using a Convolutional Neural Network (CNN).

Multiple fake classes were used to encourage early convergence and improve sample quality. Similar to FSC-GAN[1], multiple fake classes were prepared from the binary facial attributes in the dataset CelebA. Real classes/labels are the attributes from the original training data, and the fake classes/labels are the associated classes/labels of images obtained during the generator training. In the same respect, a fake facial attribute label for a generated image is obtained by doubling the size of the label embedding. For instance, the binary attribute eyeglasses represented as 0100 instead of 01 and the associated fake eyeglasses label would be 0001.

Our GAN model has an auxiliary classifier and is trained using a modified AC-GAN objective. The discriminator object maximizes the sampling loss and the sum of the classification loss over real samples and fake samples into the corresponding real or fake classes as shown in Equation4 below. The generator maximizes the difference between sampling and the classification loss of real and fake samples into real classes only as shown in Equation 5.

$$\begin{aligned}\mathcal{L}_{\mathrm{s}} = \mathbb{E}[\log P\left(S = real|X_{real}\right)] \\ +\mathbb{E}[\log P\left(S = fake|X_{fake}\right)]\end{aligned} \tag{1}$$

$$\begin{aligned}\mathcal{L}_{\mathrm{cd}} = \mathbb{E}[\log P\left(C = c|X_{real}\right)] \\ +\mathbb{E}[\log P\left(C' = c'|X_{fake}\right)]\end{aligned} \tag{2}$$

$$\begin{aligned}\mathcal{L}_{\mathrm{cg}} = \mathbb{E}[\log P\left(C = c|X_{real}\right)] \\ +\mathbb{E}[\log P\left(C = c|X_{fake}\right)]\end{aligned} \tag{3}$$

$$\mathcal{L}_{\mathrm{D}} = \mathcal{L}_{\mathrm{s}} + \mathcal{L}_{\mathrm{cd}} \tag{4}$$

$$\mathcal{L}_{\mathrm{G}} = \mathcal{L}_{\mathrm{s}} - \mathcal{L}_{\mathrm{cg}} \tag{5}$$

Where X_{real} and X_{fake} are the set of real training data and generated images respectively, C represent real facial attributes and C' represents the associated fake labels. $\mathcal{L}_{\mathrm{s}}$ is the sampling loss which represents the probability of an image being real or fake, $\mathcal{L}_{\mathrm{D}}$ is the discriminator loss, and $\mathcal{L}_{\mathrm{G}}$ is the generator loss. Our training procedure employs oversampling to emphasize equal participation of the minority classes. Algorithm 1 summarises the training procedure of MFC-GAN.

Algorithm 1 MFC GAN Training procedure

```
for i in iterations do
    mini batch ← next training batch
    evaluate L_D using mini batch
    evaluate L_G using mini batch
    if i in steps then
        for j < k_ministeps do
            mini batch ← next minority batch
            evaluate L_D using mini batch
            evaluate L_G using mini batch
        end for
    end if
end for
```

Both *steps* and *mini steps* are hyper-parameters which are tunable, and they control the behaviour of the oversampling routine. For this experiment, the *steps* variable was kept at a value of 1000 and a *mini steps* of 50 was used. The generator model has one linear layer and five transpose convolution layers with strides of two in each layer. Batch normalization was used between adjacent layers and all layers were activated using LeakyReLu apart from the final layer that is sigmoid

activated. The generator takes as input a random noise vector and the facial attributes as embeddings. The output is a 64×64 coloured image that is sent to the discriminator for training. The discriminator is trained on two set of images, the real training samples, and the generated samples. The first four layers are convolution layers with strides of two which are activated using LeakyReLU and batch normalization is used between layers. The final layer is parallel linear layer sigmoid output and a classification layer. We used a batch size of 100 and a learning rate of $1e-4$. Spectral normalisation [18] was used in both the generator and the discriminator, and we also experimented with gradient penalty [9]. Figure 1 presents a schematic overview of our model.

IV. Experiment

Our experiments analyze class-specific image generation and classification in a class imbalanced dataset. Experiments were conducted on celebrity faces with attributes dataset (CelebA dataset). We considered two facial attributes namely; eyeglasses and Goatee as minority classes with 13193 and 12716 instances respectively. Different experiments were carried out on a reduced number of instances using these minority classes. We considered 200, 500, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000 and 10000 minority instances in different runs. Setting the number of minority class instances to range between 200 to 10000 allow us to assess our model in generating data in extreme imbalanced scenarios where the majority class instances represent almost 95% to 99% of the data, while the minority class instances presence ranges roughly between 0.1% to 5% of the data.

For our image generation experiment, we report the quality and diversity of the generated minority samples after each run. For classification experiments, we extend the training data with generated minority samples from trained models (AC-GAN and MFC-GAN). Then, a CNN classifier is trained on the extended dataset, and the classification performances on the minority classes are reported.

A. Dataset

CelebA was created by annotating images from CelebFaces dataset with a face bounding box, facial landmarks and attributes annotations. It consists of 202k images with forty binary facial attributes. CelebA dataset is used as a benchmark in face detection and facial landmarks detection such as eyes, nose and mouth and facial attribute classification. CelebA attributes include curly hair, goatee, bald, male, eyeglasses, and other fine-grained attributes like wearing lipstick, heavy make-up, 5 O'clock shadow, arched eyebrows, and others. The multi-label attributes of an image open some interesting scenarios when investigating the dataset. These include the relationship between some attributes such as young and attractive, the biased distribution of attributes across samples and an unconstrained environment in facial images which creates variation among similar attributes. For our experiments, the dataset was used to perform face generation and classification of facial attributes using a low number of instances of a particular class. The dataset was preprocessed by cropping the head region using the face annotation bounding box and some heuristics. The crop was made just enough to accommodate the chin to the hair with ears visible on both sides (where applicable). The cropped image is then resized to 64×64 image patch. Before training, the images were normalized and labels preprocessed as described in section III. The dataset was split into a train and a test set. The test set is made up of six thousand samples with an equal number of majority and minority samples.

B. Face Generation from Attributes

Control generation was achieved by conditioning the generator on attribute labels. Several experiments were carried out with a different number of samples in the minority classes specifically eyeglasses and goatee classes. For each run, the MFC-GAN model is trained from the scratch and samples are generated after the training is completed. A similar experiment was performed with AC-GAN using goatee and eyeglasses attributes. For a fair comparison, a similar generator and discriminator structure was used in AC-GAN. We then examine the quality of the generated samples and how suitable these samples are for augmentation. Samples are good enough if they are of high quality and the required minority attribute appears in the image. The quality of the generated images from the two models is compared using established qualitative measures. We employ visual inspection and Frechet Inception Distance (FID) [12] to evaluate the quality and diversity of MFC-GAN and AC-GAN samples. A lower FID indicates a better sample quality and diversity. Visual inspection reaffirms the presence or absence of the attribute in a generated sample.

C. Facial Attributes Classification

Our classification model is a CNN with the same structure as the attribute CNN [15]. The attribute CNN has four convolution layers with max pooling layers between them. A fully connected layer follows the last convolution layer with a classifier as the final layer. We used a soft-max classifier, a filter size of three by three in all layers and trained the CNN from scratch as against starting from pre-trained weight as in [15]. We performed an initial classification of samples using reduced number of samples in the minority classes (eyeglasses and goatee). We refer to this experiment as a baseline. The number of samples in the minority classes is then extended with MFC-GAN generated samples after training on the same number of minority samples, and the classifier is retrained again. In a similar manner, AC-GAN samples were also used to extend the training data, and the CNN is trained from scratch. Finally, we report the F1-score and True positive rate of the classifier on each run. We compare the performances of the CNN when MFC-GAN samples are added to when AC-GAN samples are added.

V. Results

The experiments conducted were used to evaluate the performance of the models on face generation and CNN classification with a reduced number of samples in the minority classes.

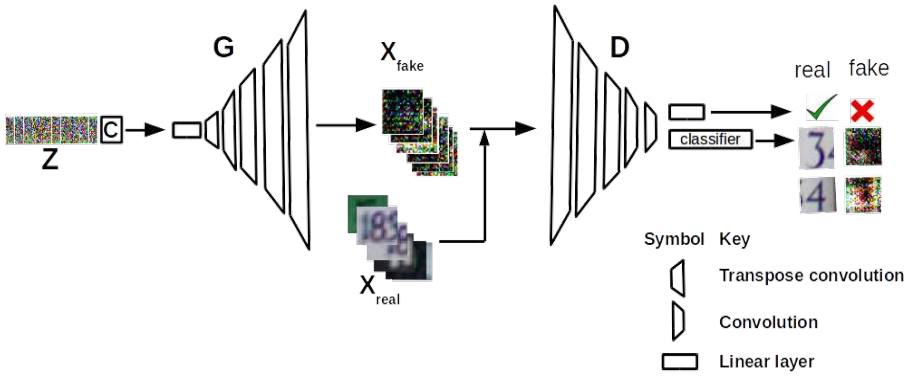


Fig. 1: *An overview of MFC-GAN model with generator network G, and discriminator network D. C is the label embedding, and X_{real} and X_{fake} are the set of real and generated samples respectively.*

Hence, we trained AC-GAN & MFC-GAN on the dataset with the varying number of samples in the minority class. Each model was then used as a source of augmentation and the classifier is retrained for each experiment. Figure 2 shows the sample data generated from each model by conditioning on the minority attribute. Tables III and IV analyse further the quality of generated images obtained during the experiments using FID metric. The FID was measured by comparing 10k samples with eyeglasses/goatee from the training data and generated 10k samples from the models after training using the approach provided in [12]. The classification results on the test set are shown in Figure 3 which compares the performance of the baseline classifier and the two models using a varying number of minority class instances. Tables I and II show the true positive rates (TPR) obtained when the models are used to augment the original dataset with more samples (generated) in a classification task. These results show clearly that when augmenting the dataset with MFC-GAN generated samples, the TPR was significantly improved in comparison with the baseline, particularly in extreme imbalanced cases (i.e. with 200 to 2000 samples). The results also show that MFC-GAN significantly outperformed AC-GAN in all scenarios.

VI. Discussion

As can be seen in the results, the CNN classifier fails to detect minority class instances in extreme scenarios. This is evident from the results shown in tables I, II & F1-scores from Figure 3. However, as the number of minority class instances increases, the CNN performance tends to improve slightly. A reasonable performance was obtained by the CNN when the number of minority class instances reached 3k for both eyeglasses and goatee attributes. This clearly shows that in extreme cases where the number of minority class instances is minimal, a data augmentation is much needed and MFC-GAN samples becomes useful.

Eyeglasses is a more prominent attribute when compared to goatee and as such better image quality and classification performance was obtained on eyeglasses attribute. Generally, both AC-GAN and MFC-GAN generated realistic samples, however, AC-GAN samples did not fall within the required

TABLE I: *True positive rate of eyeglass attribute classification and highlighted in bold are the instances where MFC-GAN performed better than the baseline and AC-GAN.*

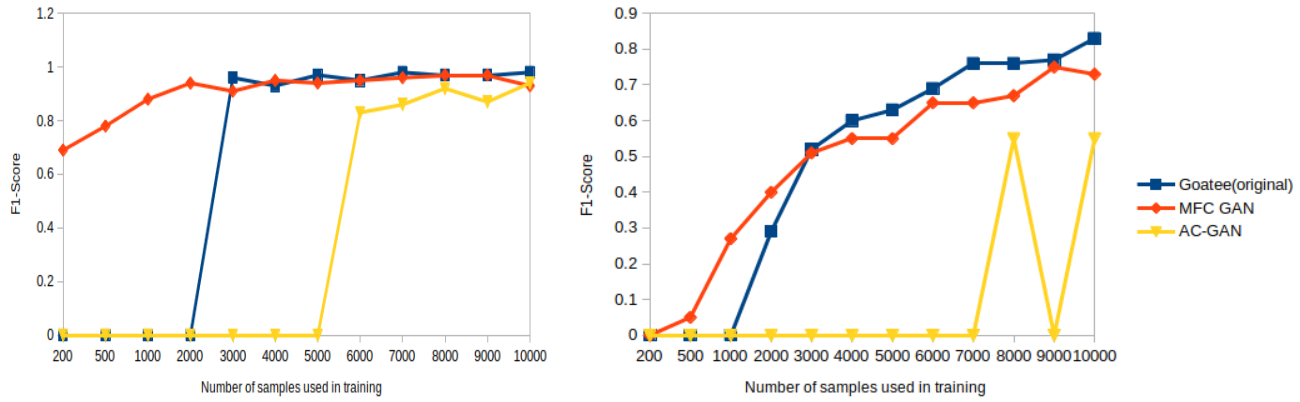
Number of samples	Baseline	AC-GAN	MFC-GAN
200	0.0	0.0	**0.53**
500	0.0	0.0	**0.64**
1000	0.0	0.0	**0.79**
2000	0.0	0.0	**0.90**
3000	0.92	0.0	0.84
4000	0.86	0.0	**0.91**
5000	0.95	0.0	0.89
6000	0.91	0.71	**0.91**
7000	0.96	0.75	0.92
8000	0.94	0.86	**0.95**
9000	0.95	0.78	0.94
10000	0.95	0.89	0.93

TABLE II: *True positive rate report on goatee attribute classification and highlighted in bold are the instances where MFC-GAN performed better than both the baseline and AC-GAN.*

Number of samples	Baseline	AC-GAN	MFC-GAN
200	0.0	0.0	0.0
500	0.0	0.0	**0.03**
1000	0.0	0.0	**0.16**
2000	0.17	0.0	**0.25**
3000	0.35	0.0	0.34
4000	0.44	0.0	0.38
5000	0.47	0.0	0.39
6000	0.53	0.0	0.49
7000	0.62	0.0	0.49
8000	0.62	0.38	0.51
9000	0.63	0.0	0.60
10000	0.72	0.38	0.58

(a) Original sample (b) AC-GAN sample (c) MFC-GAN sample

Fig. 2: *Samples were generated after the training was completed. The first row shows samples with eyeglasses attributes and the second row shows samples with goatee attribute. The leftmost are the original eyeglasses samples, in the middle are AC-GAN generated samples and the third column shows the samples generated using MFC-GAN. Samples were generated from training on 10k glasses/goatee instances.*



(a) *Eyeglasses classification performance* (b) *Goatee classification performance*

Fig. 3: *The Figure shows the classification performance over varying number of samples in the minority classes and using AC-GAN & MFC-GAN as the augmentation samples sources. The left Figure shows the F1-score of the CNN classifier over reduced eyeglasses attribute and the right hand side Figure shows the F1-score of the classifier over reduced goatee attributes.*

TABLE III: *Mean Freschet Inception Distance (FID) of generated images from experiments on different number of samples with eyeglasses.*

Number of samples	AC-GAN	MFC-GAN
200	72.97	81.26
500	73.51	**72.54**
1000	72.24	**69.65**
2000	75.34	83.36
3000	75.58	81.83
4000	73.18	**65.31**
5000	74.66	**68.02**
6000	75.38	**60.19**
7000	71.90	**59.64**
8000	70.67	**70.57**
9000	74.69	**57.17**
10000	73.96	**59.34**

TABLE IV: *Mean Freschet Inception Distance (FID) of generated image from experiments on different number of samples with goatee.*

Number of samples	AC-GAN	MFC-GAN
200	71.45	**65.64**
500	68.15	**65.80**
1000	69.14	**69.67**
2000	66.62	**66.79**
3000	58.72	62.56
4000	69.48	**62.16**
5000	57.06	61.25
6000	70.55	71.35
7000	92.97	**61.37**
8000	73.29	**68.51**
9000	60.02	**59.69**
10000	70.44	**61.69**

category whereas MFC-GAN model performs significantly better when the dataset was extremely imbalanced. Moreover, the samples from MFC-GAN proved to be useful in augmenting the training set to boost classification performance. For instance, with 200 eyeglasses samples, the true positive and F1-score improved from 0 to 53 and $\sim$70% respectively.

Augmenting AC-GAN samples into the training set did not improve the classification accuracy of the CNN in both eyeglasses and goatee attributes classification. Moreover, the worse results were obtained in goatee classification. Visual observation of the generated images revealed that the model generated quality samples but was not of the required minority classes. Adding these samples to the training data confuses the classification model more particularly when the number of original samples is small. These samples over-shadow the real data and prevented the model from understanding the true discriminative feature/attribute in the samples. However, with significant real samples in the training data, the effect of the spurious samples is minimised. This is in line with the observations by [16] and shows that AC-GAN is inadequate in capturing the true data distribution in an extreme class imbalance scenario.

Augmenting MFC-GAN in the training set resulted in better classification results. This improvement was significant in extreme cases where the number of minority class instances is kept to minimal. MFC-GAN model trained on two hundred eyeglasses samples was able to capture the real data distribution and was capable of producing the required minority samples necessary to improve classification results. Visually observing the samples in Figure 2 shows the presence of the minority attributes which further explain the improvement in performance. These samples also had better mean FID than the samples generated by AC-GAN as shown in table III and IV. An interesting behaviour of the MFC-GAN model is that it was able to associate goatee with only male faces despite training on female examples while the model generated both male and female samples for eyeglasses.

Despite improving classification performance on a reduced number of samples (minority classes), we observed that augmenting more samples could not achieve 100% true positive rate even with 10k real samples. We tried to push the results further by under-sampling the majority class but this did not influence the results much. We infer that this could be related to the classification model chosen because no hyper-parameter search or model tuning was done. In addition, the target of our experiments was to show the usefulness of our GAN generated samples in extreme class imbalanced scenarios and classification was only used as an evaluation criterion.

VII. Conclusion

In this paper, we presented a Multiple Fake Classes Generative Adversarial Networks (MFC-GAN) to generate face images from an under-represented class in the dataset. We applied our method to a face generation task conditioned on facial attributes. Several experiments were carried out on a reduced number of instances in the classes of interest. Results obtained showed that MFC-GAN was able to capture the underlying data distribution from a class imbalanced dataset and generated realistic samples from a minority class instances. Furthermore, MFC-GAN samples were used to improve attribute classification in the minority classes through augmentation. The results obtained showed that MFC-GAN improved the baseline classification in extreme imbalance scenario while out-performing AC-GAN in all cases.

In our future work, we will study the relationship between attributes and how this relationship affects the multi-class imbalanced problem. Some of these facial attributes occur consistently alongside each other such as male and goatee, attractive and young. Others such as beard and sideburns or beard and moustache frequently occur together in the dataset but are independent of one another. Trying to improve the number of samples in such classes using sample generation

may indirectly affect the other. Exploring how an augmentation model will isolate and maintain a balance between these subtle attributes will be an interesting research area.

REFERENCES

[1] Ali-Gombe Adamu, Elyan Eyad, Savoye Yann, and Jayne Chrisina. Few-shot classifier gan. In *Neural Networks (IJCNN), 2018 International Joint Conference on*. IEEE, 2018.

[2] Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017.

[3] Christoph Baur, Shadi Albarqouni, and Nassir Navab. Melanogans: High resolution skin lesion synthesis with gans. *arXiv preprint arXiv:1804.04338*, 2018.

[4] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

[5] Max Ehrlich, Timothy J Shields, Timur Almaev, and Mohamed R Amer. Facial attributes classification using multi-task representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 47–55, 2016.

[6] Maayan Frid-Adar, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Synthetic data augmentation using gan for improved liver lesion classification. *arXiv preprint arXiv:1801.02385*, 2018.

[7] Jon Gauthier. Conditional generative adversarial nets for convolutional face generation. *Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester*, 2014(5):2, 2014.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[9] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017.

[10] Emily M Hand, Carlos D Castillo, and Rama Chellappa. Doing the best we can with what we have: Multi-label balancing with selective learning for attribute prediction. In *AAAI*, 2018.

[11] Emily M Hand and Rama Chellappa. Attributes for improved attributes: A multi-task network utilizing implicit and explicit relationships for facial attribute classification. In *AAAI*, pages 4068–4074, 2017.

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

[13] Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016.

[14] Neeraj Kumar, Alexander Berg, Peter N Belhumeur, and Shree Nayar. Describable visual attributes for face verification and image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):1962–1977, 2011.

[15] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.

[16] Giovanni Mariani, Florian Scheidegger, Roxana Istrate, Costas Bekas, and Cristiano Malossi. Bagan: Data augmentation with balancing gan. *arXiv preprint arXiv:1803.09655*, 2018.

[17] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[18] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957 and ICLR2018*, 2018.

[19] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. *International conference on machine learning,page 2642-2651*, 70:2642–2651, AUG 2017.

[20] Guim Perarnau, Joost van de Weijer, Bogdan Raducanu, and Jose M Álvarez. Invertible conditional gans for image editing. *arXiv preprint arXiv:1611.06355*, 2016.

[21] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[22] Xiangbo Shu, Yunfei Cai, Liu Yang, Liyan Zhang, and Jinhui Tang. Computational face reader based on facial attribute estimation. *Neurocomputing*, 236:153–163, 2017.

[23] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, pages 3630–3638, 2016.

[24] Lipeng Wan, Jun Wan, Yi Jin, Zichang Tan, Stan Z Li, et al. Fine-grained multi-attribute adversarial learning for face generation of age, gender and ethnicity, 2018.

[25] Kaipeng Zhang, Lianzhi Tan, Zhifeng Li, and Yu Qiao. Gender and smile classification using deep convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 34–38, 2016.

[26] Jingjing Zheng, Zhuolin Jiang, Rama Chellappa, and Jonathon P Phillips. Submodular attribute selection for action recognition in video. In *Advances in Neural Information Processing Systems*, pages 1341–1349, 2014.