Multiple fake classes GAN for data augmentation in face image dataset.

ALI-GOMBE, A., ELYAN, E. and JAYNE, C.

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

Face images generation and face image attributes classification are essential area in computer vision. These are used in face verification systems [11], face reading [18] and activity recognition [21]. These systems train a model on a carefully compiled training data and replicate a similar performance on a real problem. However, the diversity of the images collected (such as background, pose, scale and others) and class imbalance hinder the models from achieving this objective. Class imbalance problem is a common occurrence in security, health and banking domain. This occurs when the different classes are not equally represented in a dataset.

Class imbalance problem could be addressed at the data level to ensure equal representation of classes during training. Under-sampling and oversampling are the popular data re-sampling techniques used to rebalanced the dataset at data level. While these two have been successfully applied, under-sampling may discard important points while oversampling is prone to over-fitting [2]. Algorithm approaches adjust learning objectives to account for different losses during training. However, deciding on the cost matrix can be tricky in real application [10].

Data augmentation is a commonly used technique in classification that reduces the effect of class imbalance problem at the data level. Augmentation improves generalisation by creating more samples needed in rebalancing the classes. 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 [17]. 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. Generative Adversarial Networks (GANs) are state-of-the-art in image generation and Conditional GAN (C-GAN) [14] 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) [16], C-GAN and GAN [5] avoid generating minority classes in extreme class imbalance cases [13].

In this paper, we propose a novel Multiple Fake Classes GAN model (MFC-GAN) for performing image generation from imbalance classes. 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, 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. CelebA attributes are represented as binary labels describing the presence or absence of a face feature such as beard or 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 in these classes. 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 papers are as follows:

- Multiple Fake Classes GAN model (MFC-GAN) for generating additional samples of face images with specific facial attributes in extreme imbalance scenario
- Apply the generated samples to rebalance the dataset and improve classification performance in relation to the under represented attributes

The remainder of this paper is organised 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 facial attributes vary significantly from one person to another [3]. Face pose angles, different lighting conditions and variety of clothing such as eyeglasses, caps and jewellery 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 on outputs from attribute detectors. The problem with this group is the sole reliance on the efficiency of the detection model [3]. And the second approaches process the global image to extract the required features and classify the attributes. The latter approaches are more robust and have provided state-of-the-art performances recently. Furthermore, global approaches have been implemented as multi-tasking approaches [3] and in some cases employing specific models to classify each attribute [12], [20]. More recently, multi-task models have explored the correlation between attributes to improve classification performances such as [8].

Multi-task approaches are mostly multi-model and utilise shared information from related problems. For instance, [12] proposed the use of a multi-model framework to perform attributes classification. The framework consists of a face localisation network (LNet) and an attribute classification network (ANet) that feeds on a localised 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 attributes like smile and

gender. Zhang *et al.* [20] 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 feature to generate faces with specific attributes. For instance, Jon in [4] 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. Further 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) [19] 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 conditional reconstruction of the embeddings and considered three sets of attributes: age, gender and ethnicity. FM-GAN was trained on CelebA dataset and the synthesised images that were used to augment MORPH II dataset. The new dataset was evaluated using a CNN and results obtained showed that the classifier performs better when the synthetic samples were added.

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.* [7] 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 towards frontal faces, smiling and mostly young celebrity pictures. The authors argued that models trained on this dataset without putting into account such biases may perform poorly on a different domain. And balancing by re-sampling a class simply affect other class distribution as well [7].

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. The generated minority samples are then used to augment he original training data to rebalance the dataset. Finally, we validate the approach on classification task using a 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 maximises 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 maximises the difference between sampling and the classification loss of real and fake samples into real classes only as shown in equation 5.

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

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

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

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

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

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

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 normalisation 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

---

Algorithm 1 MFC GAN Training procedure

> **for** $i < 50000$ **do**
>   *mini batch* ← *next training batch*
>   evaluate $\mathcal{L}_{D}$ using *mini batch*
>   evaluate $\mathcal{L}_{G}$ using *mini batch*
>   **if** $i\%steps = 0$ **then**
>   *oversample:*
>   **for** $j < k_{ministeps}$ **do**
>   *mini batch* ← *next minority batch*
>   evaluate $\mathcal{L}_{D}$ using *mini batch*
>   evaluate $\mathcal{L}_{G}$ using *mini batch*
>   **end for**
>   **end if**
> **end for**

---

training samples and the generated samples. The first four layers are convolution layers with strides of two which are activated using LeakyRelU and batch normalisation is used between layers. The final layer is parallel linear layer sigmoid output and a classification layer. Spectral normalisation [15] was used in both the generator and the discriminator, and we also experimented with gradient penalty [6].

IV. EXPERIMENT

Our experiments analyse class-specific image generation and classification in a class imbalance 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 in this minority classes. We considered 200, 500, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000 and 10000 minority instances in different runs. 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 includes 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 low instances in a 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. Figures 2 and 1 shows the distribution of attributes in CelebA dataset and sample images respectively after preprocessing is complete. Prior to training, the images are normalised 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.
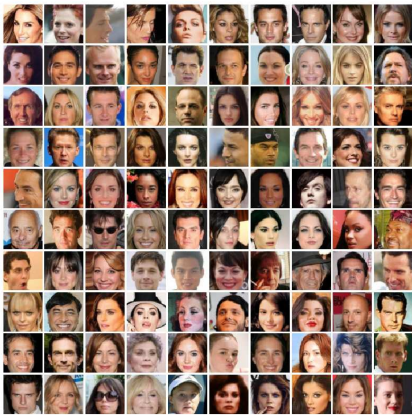

Fig. 1: *Samples images from CelebA dataset after preprocessing.*

B. Face Generation from Attributes

Control generation was achieved by conditioning the generator on attribute labels. Several experiments were carried out with 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. 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) [9] 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. The CNN used has the same structure as the attribute CNN [12]. 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 softmax classifier, a filter size of three in all layers and trained the CNN from scratch as against starting from pre-trained weight as in [12]. 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 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. RESULT

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. Hence, we trained AC-GAN & MFC-GAN on the dataset with varying number of samples in the minority class. Each model was then used as a source of augmentation samples and the classifier is retrained for each experiment. Figure 3 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 measure 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 [9]. The classification results on the test set are reported in figure 4 which compares the performance of the baseline classifier and the two models using varying number of minority classes. Tables I and II show the true positive rates obtained when the models are used as augmentation approaches in a classification task. MFC-GAN samples improved the baseline true positive rates by rebalancing the dataset particularly in extreme cases while outperforming AC-GAN in all cases.

VI. DISCUSSION

When the CNN classifier is trained on reduced number of samples without augmentation, the classifier struggles to identify any of the minority samples used in the test set. This behaviour is observed from the true positive rates in tables I, II & f1-scores from figure 4. However, as the number of minority samples increases, the performance tends to improve. Reasonable performance was by the CNN (without augmentation) was recorded when the number of samples reaches 3k and for both eyeglasses and goatee attribute. This clearly indicates that with extreme imbalance classes, training directly on the dataset yields undesirable results.

Eyeglasses is a more prominent attribute when compared to goatee and as such better image quality and classification
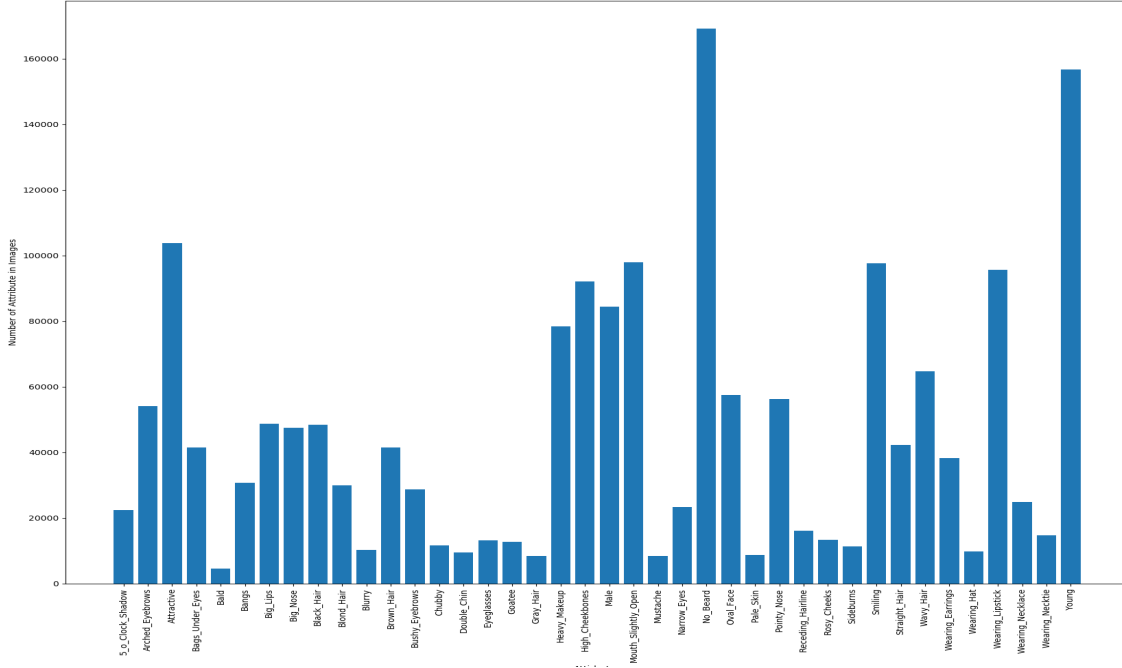
Fig. 2: *The number of samples per attribute class in CelebA dataset.*

TABLE I: *True positive rate report on eyeglasses classification*

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 classification*

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

performance was obtained on eyeglasses. Generally, both AC-GAN and MFC-GAN generated realistic samples however, AC-GAN samples did not fall within the required category. Whereas MFC-GAN model was able to train on imbalance classes even in extreme conditions. Moreover, the samples from MFC-GAN proved to be useful in augmenting the training samples to boost classification performance. For instance, with 200 eyeglasses samples, the true positive and f1-score improved from 0 to 53 and ∼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 were 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

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

Fig. 3: *Samples were generated after the training is completed. The first row are samples with eyeglasses attributes and the second row are samples with goatee attribute. The left most are the original eyeglasses samples, in the middle are AC-GAN generated samples and the third column are the samples generated using MFC-GAN. Samples are generated from training on 10k glasses/goatee instances.*
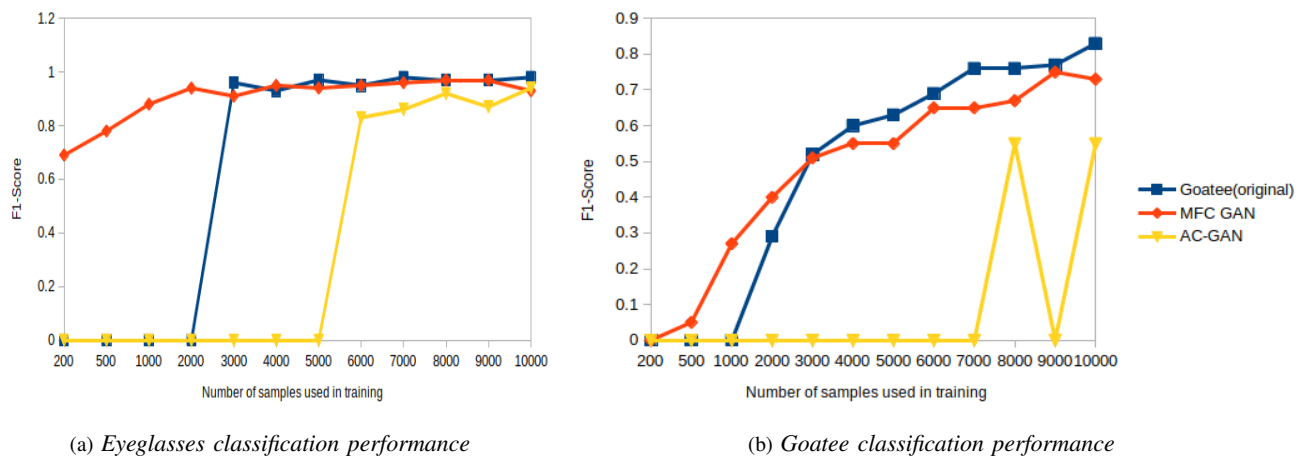

(a) *Eyeglasses classification performance*   (b) *Goatee classification performance*

Fig. 4: *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**

observations by [13] and shows that AC-GAN is inadequate in capturing the true data distribution in an extreme class imbalance scenario.

Using generated samples from MFC-GAN for augmentation, better classification performances were recorded especially in extreme conditions for both eyeglasses and goatee experiments. MFC-GAN model trained on two hundred eyeglasses samples was able to capture the true data distribution and produced the required minority samples necessary to improve classification results. Visually observing the samples in figure 3 shows the presence of the minority attributes which further explain the improvement in performance. These samples also hard 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 with eyeglasses.

Despite improving classification performance on reduced number of samples, 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 tuning model was done. In addition, the target of our experiments was to show the usefulness of our GAN generated samples and classification was used as an evaluation criterion.

VII. CONCLUSION

In this paper, we proposed a Multiple Fake Classes Generative Adversarial Networks (MFC-GAN) that generate samples from few instances in a class. We apply MFC-GAN on 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 imbalance dataset and generated realistic samples from the required minority classes. 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 multi-class imbalance 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] 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.
[3] 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.
[4] 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.
[5] 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.
[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

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

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

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

[16] 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.

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

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

[19] 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.

[20] 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.

[21] 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.