

Generative Adversarial Minority Oversampling

Sankha Subhra Mullick
Indian Statistical Institute
Kolkata, India
sankha_r@isical.ac.in

Shounak Datta
Duke University
Durham, NC, USA
shounak.jaduniv@gmail.com

Swagatam Das
Indian Statistical Institute
Kolkata, India
swagatam.das@isical.ac.in

Abstract

Class imbalance is a long-standing problem relevant to a number of real-world applications of deep learning. Oversampling techniques, which are effective for handling class imbalance in classical learning systems, can not be directly applied to end-to-end deep learning systems. We propose a three-player adversarial game between a convex generator, a multi-class classifier network, and a real/fake discriminator to perform oversampling in deep learning systems. The convex generator generates new samples from the minority classes as convex combinations of existing instances, aiming to fool both the discriminator as well as the classifier into misclassifying the generated samples. Consequently, the artificial samples are generated at critical locations near the peripheries of the classes. This, in turn, adjusts the classifier induced boundaries in a way which is more likely to reduce misclassification from the minority classes. Extensive experiments on multiple class imbalanced image datasets establish the efficacy of our proposal.

1. Introduction

The problem of class imbalance occurs when all the classes present in a dataset do not have equal number of representative training instances [19, 11]. Most of the existing learning algorithms produce inductive bias favoring the majority class in presence of class imbalance in the training set, resulting in poor performance on the minority class(es). This is a problem which routinely plagues many real-world applications such as fraud detection, dense object detection [30], medical diagnosis, etc. For example, in a medical diagnosis application, information about unfit patients is scarce compared to that of fit individuals. Hence, traditional classifiers may misclassify some unfit patients as being fit, having catastrophic implications [32].

Over the years, the machine learning community has devised many methods for tackling class imbalance [24, 4]. However, only a few of these techniques have been extended to deep learning even though class imbalance is

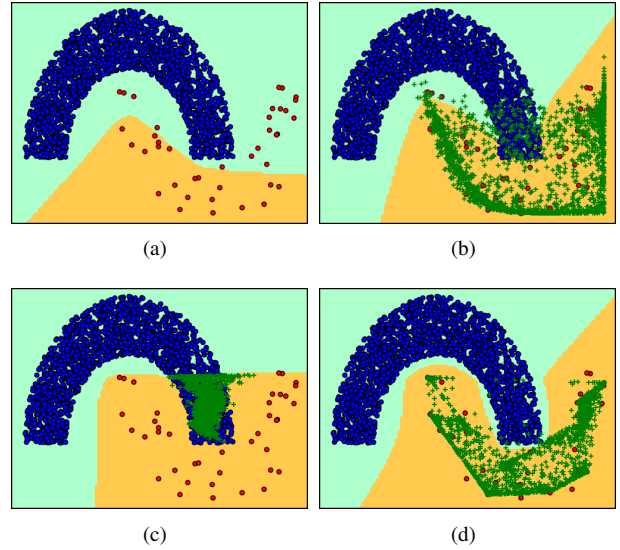


Figure 1. Illustration using a ‘toy’ dataset: (a) Imbalanced classification with an unaided classifier network M results in misclassification of the minority class instances (red dots). (b) Artificial minority points (green ‘+’) generated using conditional GAN help to improve the result on the minority class but bleed into the majority class (blue dots), affecting the performance on the latter. (c) New points are generated by training a convex generator G alternately with M . This is a two player adversarial game where G attempts to generate samples which are hard for M to correctly classify. This results in ideal performance on the minority class, but at the cost of misclassifying the majority class as G does not adhere to the distribution of the minority class. (d) Ideal performance on both classes is achieved by further incorporating an additional discriminator D to induce fidelity to the minority class distribution and to limit bleeding into majority class territory.

fairly persistent in such networks, severely affecting both the feature extraction as well as the classification process [48, 21, 49, 5, 22]. The existing solutions [21, 10, 43, 30, 6] for handling class imbalance in deep neural networks mostly focus on cost tuning to assign suitably higher costs to minority instances. Another interesting class of approaches [50, 12] focuses on constructing balanced sub-

samples of the dataset. Wang et al. [44] proposed a novel meta-learning scheme for imbalanced classification. It is interesting to note that oversampling techniques like SMOTE [8] have not received much attention in the context of deep learning, despite being very effective for classical systems [14]. This is because deep feature extraction and classification are performed in an end-to-end fashion, making it hard to incorporate oversampling which is typically done subsequent to feature extraction. An attempt to bridge this gap was made by Ando and Huang [1] in their proposed deep oversampling framework (DOS). However, DOS uniformly oversamples the entire minority class and is not capable of concentrating the artificial instances in difficult regions. Additionally, the performance of DOS depends on the choice of the class-wise neighborhood sizes, which must be determined by costly parameter tuning.

Generative adversarial networks (GANs) are a powerful subclass of generative models that have been successfully applied to image generation. This is due to their capability to learn a mapping between a low-dimensional latent space and a complex distribution of interest, such as natural images [15, 33, 36, 35]. The approach is based on an adversarial game between a generator that tries to generate samples which are similar to real samples and a discriminator that tries to discriminate between real training samples and generated samples. The success of GANs as generative models has led Douzas and Bacao [13] to investigate the utility of using GANs to oversample the minority class(es). However, attempting to oversample the minority class(es) using GANs can lead to boundary distortion [39], resulting in a worse performance on the majority class (as illustrated in Figure 1(b)). Moreover, the generated points are likely to lie near the mode(s) of the minority class(es) [42], while new points around the class boundaries are required for learning reliable discriminative (classification) models [17, 18].

Hence, in this article, we propose (in Section 3) a novel end-to-end feature-extraction-classification framework called Generative Adversarial Minority Oversampling (GAMO) which employs adversarial oversampling of the minority class(es) to mitigate the effects of class imbalance. The contributions made in this article differ from the existing literature in the following ways:

1. Unlike existing deep oversampling schemes [1, 13], GAMO is characterized by a three-player adversarial game among a convex generator G , a classifier network M , and a discriminator D .
2. Our approach is fundamentally different from existing adversarial classification schemes (where the generator works in harmony with the classifier to fool the discriminator) [38, 27, 41, 35], in that our convex generator G attempts to fool both M and D .
3. Unlike the generator employed in GAN [15], we constrain G to conjure points within the convex hull of

the class of interest. Additionally, the discriminator D further ensures that G adheres to the class distribution for non-convex classes. Consequently, the adversarial contention with M pushes the conditional distribution(s) learned by G towards the periphery of the respective class(es), thus helping compensate for class imbalance effectively.

4. In contrast to methods like [8, 13], G can oversample different localities of the data distribution to different extents based on the gradients obtained from M .
5. For applications requiring a balanced training set of images, we also propose a technique called GAMO2pix (Section 5) that can generate realistic images from the synthetic instances generated by GAMO in the distributed representation space.

We undertake an ablation study as well as evaluate the performance of our method compared to the state-of-the-art in Section 4, and make concluding remarks in Section 6.

2. Related Works

The success of SMOTE [8, 9] has inspired several improvements. For example, [17, 7] attempt to selectively oversample minority class points lying close to the class boundaries. Works like [18, 29, 2], on the other hand, asymmetrically oversample the minority class such that more synthetic points are generated surrounding the instances which are difficult to classify. Although these methods achieved commendable improvement on classical classifiers, they can neither be extended to deep learning techniques nor be applied to images, respectively due to the end-to-end structure of deep learning algorithms and a lack of proper notion of distance between images.

Extending GANs for semi-supervised learning, works like [27, 38] fused a c -class classifier with the discriminator by introducing an extra output line to identify the fake samples. On the other hand, [41] proposed a c -class discriminator which makes uncertain predictions for fake images. Additionally, [35] proposed a shared discriminator-cum-classifier network which makes two separate sets of predictions using two different output layers. These approaches can loosely be considered to be related to GAMO as these also incorporate a classifier into the adversarial learning scheme.

3. Proposed Method

Let us consider a c -class classification problem with a training dataset $X \subset \mathbb{R}^D$ (of images vectorized either by flattening or by a convolutional feature extraction network F). Let the prior probability of the i -th class be P_i , where $i \in \mathcal{C} = \{1, 2, \dots, c\}$; \mathcal{C} being the set of possible class labels. Without loss of generality, we consider the classes to be ordered such that $P_1 \leq P_2 \leq \dots < P_c$. Moreover, let X_i

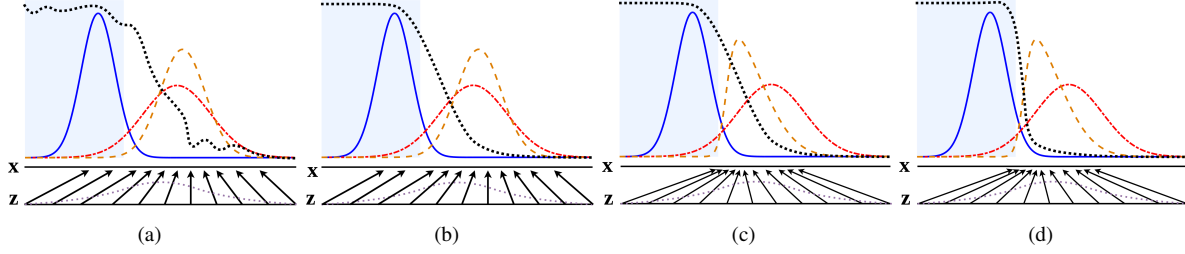


Figure 2. GAMO functions by simultaneously updating the classifier M and the generator G . The classification function (black, dotted line) is trained to correctly classify samples from the majority class distribution p_{maj}^d (blue, solid line), the real minority class distribution p_{min}^d (red, dots and dashes) as well as the generated minority distribution p_{min}^g (brown, dashed line). The generator, on the other hand, is trained to generate minority samples which will be misclassified by M . The upward arrows show how the generator learns the mapping $\mathbf{x} = G(\mathbf{z})$ from a standard normal distribution (mauve, dotted line) in the latent space to convex combinations of the real minority instances from the minority class. The ideal classification function is shown as a blue highlight in the background. (a) Let us consider an initial adversarial pair: the generated distribution p_{min}^g is similar to the real distribution of the minority class p_{min}^d and M is an inaccurate classifier. (b) M is trained to properly classify the samples from the three distributions p_{maj}^d , p_{min}^d , and p_{min}^g ; resulting in a non-ideal trained classifier which is biased in favor of the majority class. (c) After an update to G , the gradient of M has guided $G(\mathbf{z})$ to flow to regions that are more likely to be misclassified by M . (d) Thereafter, retraining M results in a classifier much closer to the ideal classifier due to the increased number of minority samples near the boundary of the two classes.

denote the set of all n_i training points which belong to class $i \in \mathcal{C}$. We intend to train a classifier M having c output lines, where the i -th output $M_i(\mathbf{x})$ predicts the probability of any $\mathbf{x} \in X$ to be a member of the i -th class.

3.1. Adversarial Oversampling

Our method plays an adversarial game between a classifier that aims to correctly classify the data points and a generator attempting to spawn artificial points which will be misclassified by the classifier. The idea is that generating such difficult points near the fringes of the minority class(es) will help the classifier to learn class boundaries which are more robust to class imbalance. In other words, the performance of the classifier will adversarially guide the generator to generate new points at those regions where the minority class under concern is prone to misclassification. Moreover, the classifier will aid the generator to adaptively determine the concentration of artificial instances required to improve the classification performance in a region, thus relieving the user from tuning the amount of oversampling. Instead, we only need to fix the number of points to be generated to the difference between the number of points in the majority class and that of the (respective) minority class(es).

3.2. Convex Generator

The generator tries to generate points which will be misclassified by the classifier. Hence, if left unchecked, the generator may eventually learn to generate points which do not coincide with the distribution of the intended minority class. This may help improve the performance on the concerned minority class but will lead to high misclassification from the other classes. To prevent this from happening, we generate the new points only as convex combinations

of the existing points from the minority class in question. This will restrict the generated distribution within the convex hull of the real samples from the (respective) minority class(es). Since the generator attempts to conjure points that are difficult for the classifier, the points are generated near the peripheries of the minority class(es).

Our convex generator G comprises of two modules: a Conditional Transient Mapping Unit ($cTMU$) and a set of class-specific Instance Generation Units (IGU), which we propose to limit the model complexity. The $cTMU$ network learns a mapping t , conditioned on class i , from a l -dimensional latent space to an intermediate space. The IGU_i , on the other hand, learns a mapping g_i from the $cTMU$ output space to a set of n_i convex weights $g_i(t(\mathbf{z}|i)) \geq 0$, s.t. $\sum_{j=1}^{n_i} g_i(t(\mathbf{z}|i)) = 1$, using softmax activation. Hence, G can generate a new D -dimensional sample for the i -th class as a convex combination of the data points in X_i ,

$$G(\mathbf{z}|i) = \sum_{j=1}^{n_i} g_i(t(\mathbf{z}|i)) \mathbf{x}_j, \quad (1)$$

where \mathbf{z} is a latent variable drawn from a standard normal distribution and $\mathbf{x}_j \in X_i$.

Formally, the adversarial game played by the proposed classifier-convex generator duo poses the following optimization problem, when cross entropy loss is considered:

$$\min_G \max_M J(G, M) = \sum_{i \in \mathcal{C}} J_i, \quad (2)$$

where $J_i = (J_{i1} + J_{i2} + J_{i3} + J_{i4})$,

$$J_{i1} = P_i \mathbb{E}_{\mathbf{x} \sim p_i^d} [\log M_i(\mathbf{x})],$$

$$J_{i2} = \sum_{j \in \mathcal{C} \setminus \{i\}} P_j \mathbb{E}_{\mathbf{x} \sim p_j^d} [\log(1 - M_i(\mathbf{x}))],$$

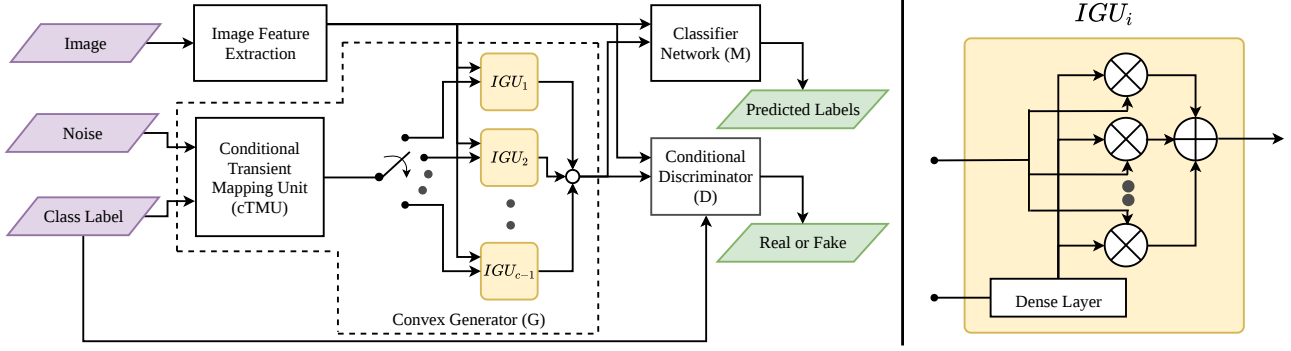


Figure 3. The GAMO model: (Left) Schematic of the GAMO framework; (Right) Illustration of an Instance Generation Unit (IGU). Given an image, the extracted feature vectors (either by a convolutional neural network F or by flattening) are fed to the classifier network M as well as the conditional discriminator D . M predicts the class label for the input data point while D distinguishes between real and fake data instances. The convex generator network G is composed of a $cTMU$, and $IGUs$ corresponding to each of the $c - 1$ minority classes. The IGU_i network takes an intermediate vector generated by $cTMU$ and maps it to a set of n_i convex weights. It then takes the set X_i as input and generates a new sample for the i -th class, as the convex combination of all the $\mathbf{x}_j \in X_i$.

$$J_{i3} = (P_c - P_i) \mathbb{E}_{G(\mathbf{z}|i) \sim p_i^g} [\log M_i(G(\mathbf{z}|i))], \text{ and,}$$

$$J_{i4} = \sum_{j \in C \setminus \{i\}} (P_c - P_j) \mathbb{E}_{G(\mathbf{z}|j) \sim p_j^g} [\log(1 - M_i(G(\mathbf{z}|j)))],$$

while p_i^d and p_i^g respectively denote the real and generated class conditional probability distributions of the i -th class.

The two-player minimax game formalized in (2) is played between a classifier M and a generator G . M attempts to correctly classify all real as well as generated points belonging to all the classes. Whereas, G strives to generate sample(s) which have a high probability of being classified by M into all other classes. To demonstrate how such an adversarial game can aid M to learn a better class boundary, we illustrate its chronological progression in a more explanatory manner in Figure 2. In Theorem 1, we show that the optimization problem in (2) is equivalent to minimizing a sum of the Jensen-Shannon divergences.

Theorem 1. *Optimizing the objective function J is equivalent to the problem of minimizing the following summation of Jensen-Shannon divergences:*

$$\sum_{i=1}^c JS\left((P_i p_i^d + (P_c - P_i) p_i^g) \parallel \sum_{\substack{j \neq i \\ j=1}}^c (P_j p_j^d + (P_c - P_j) p_j^g)\right)$$

Proof. See the supplementary document. \square

The behavior of the proposed approach can be understood by interpreting Theorem 1. The optimization problem aims to bring the generated distribution, for a particular class, closer to the generated as well as real distributions for all other classes. Since the real distributions are static for a fixed dataset, the optimization problem in Theorem 1 essentially attempts to move the generated distributions

for each class closer to the real distributions for all other classes. This is likely to result in the generation of ample points near the peripheries, which are critical to combating class imbalance. While doing so, the generated distributions for all classes also strive to come closer to each other. However, the generated distributions for the different classes do not generally collapse upon each other, being constrained to remain within the convex hulls of the respective classes.

3.3. Additional Discriminator

While the generator only generates points within the convex hull of the samples from the minority class(es), the generated points may still be placed at locations within the convex hull which do not correspond to the distribution of the intended class (recall Figure 1(c)). This is likely to happen if the intended minority class(es) are non-convex in shape. Moreover, we know from Theorem 1 that the generated distributions for different minority classes may come close to each other if the respective convex hulls overlap. To solve this problem, we introduce an additional conditional discriminator which ensures that the generated points do not fall outside the actual distribution of the intended minority class(es). Thus, the final adversarial learning system proposed by us consists of three players, viz. a multi-class classifier M , a conditional discriminator D which given a class aims to distinguish between real and generated points, and a convex generator G that attempts to generate points which, in addition to being difficult for M to correctly classify, are also mistaken by D to be real points sampled from the given dataset. The resulting three-player minimax game is formally presented in (3).

$$\min_G \max_M \max_D Q(G, M, D) = \sum_{i \in C} Q_i, \quad (3)$$

$$\text{where, } Q_i = (J_{i1} + J_{i2} + J_{i3} + J_{i4} + Q_{i1} + Q_{i2}),$$

$$Q_{i1} = P_i \mathbb{E}_{\mathbf{x} \sim p_i^d} [\log D(\mathbf{x}|i)], \text{ and,}$$

Algorithm 1 Generative Adversarial Minority Oversampling (GAMO)

Input: X : training set, l : latent dimension, b : minibatch size, u, v : (hyperparameters, set to $\lceil \frac{n}{b} \rceil$ in our implementation).

Output: A trained classification network M .

Note: For flattened images there is no need to train F , i.e., $F(X)$ can be replaced by X .

```
1: while not converged do
2:   for  $u$  steps do
3:     Sample  $B_d = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_b\}$  from  $X$ , with corresponding class labels  $Y_d$ .
4:     Update  $F$  by gradient descent on  $(M(F(B_d)), Y_d)$  keeping  $M$  fixed.
5:   end for
6:   for  $v$  steps do
7:     Sample  $B_d = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_b\}$  from  $X$ , with corresponding class labels  $Y_d$ .
8:     Sample  $B_n = \{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_b\}$  from  $l$  dimensional standard normal distribution.
9:     Update  $M$  and  $D$  by respective gradient descent on  $(M(F(B_d)), Y_d)$  and  $(D(F(B_d)|Y_d), \mathbf{1})$ , keeping  $F$  fixed.
10:    Generate labels  $Y_n$  by assigning each  $\mathbf{z}_j \in B_n$  to one of the  $c - 1$  minority classes, with probability  $\propto (P_c - P_i); \forall i \in \mathcal{C} \setminus \{c\}$ .
11:    Update  $M$  and  $D$  by respective gradient descent on  $(M(G(B_n|Y_n)), Y_n)$  and  $(D(G(B_n|Y_n)|Y_n), \mathbf{0})$ , keeping  $G$  fixed.
12:    Sample  $B_g = \{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_b\}$  from  $l$  dimensional standard normal distribution.
13:    Generate labels  $Y_g$  by assigning each  $\mathbf{z}_j \in B_g$  to any of the  $c - 1$  minority classes with equal probability. Take ones' complement of  $Y_g$  as  $\bar{Y}_g$ .
14:    Update  $G$  by gradient descent on  $(M(G(B_g|Y_g)), \bar{Y}_g)$  keeping  $M$  fixed.
15:    Update  $G$  by gradient descent on  $(D(G(B_g|Y_g)|Y_g), \mathbf{1})$  keeping  $D$  fixed.
16:   end for
17: end while
```

$$Q_{i2} = (P_c - P_i) \mathbb{E}_{G(\mathbf{z}|i) \sim p_i^g} [\log(1 - D(G(\mathbf{z}|i)|i))].$$

3.4. Least-Square Formulation

Mao et al. [31] showed that replacing the popular cross entropy loss in GAN with least square loss can not only produce better quality images but also can prevent the vanishing gradient problem to a greater extent. Therefore, we also propose a variant of GAMO using the least square loss, which poses the following optimization problem:

$$\min_M L_M = \sum_{i \in \mathcal{C}} (L_{i1} + L_{i2} + L_{i3} + L_{i4}), \quad (4)$$

$$\min_D L_D = \sum_{i \in \mathcal{C}} (L_{i5} + L_{i6}), \quad (5)$$

$$\min_G L_G = \sum_{i \in \mathcal{C} \setminus \{c\}} (L_{i7} + L_{i8} + L_{i9}), \quad (6)$$

where, $L_{i1} = P_i \mathbb{E}_{\mathbf{x} \sim p_i^d} [(1 - M_i(\mathbf{x}))^2]$,

$$L_{i2} = \sum_{j \in \mathcal{C} \setminus \{i\}} P_j \mathbb{E}_{\mathbf{x} \sim p_j^d} [(M_i(\mathbf{x}))^2],$$

$$L_{i3} = (P_c - P_i) \mathbb{E}_{G(\mathbf{z}|i) \sim p_i^g} [(1 - M_i(G(\mathbf{z}|i)))^2],$$

$$L_{i4} = \sum_{j \in \mathcal{C} \setminus \{i\}} (P_c - P_j) \mathbb{E}_{G(\mathbf{z}|j) \sim p_j^g} [(M_i(G(\mathbf{z}|j)))^2],$$

$$L_{i5} = P_i \mathbb{E}_{\mathbf{x} \sim p_i^d} [(1 - D(\mathbf{x}|i))^2],$$

$$L_{i6} = (P_c - P_i) \mathbb{E}_{G(\mathbf{z}|i) \sim p_i^g} [(D(G(\mathbf{z}|i)|i))^2],$$

$$L_{i7} = \mathbb{E}_{G(\mathbf{z}|i) \sim p_i^g} [(M_i(G(\mathbf{z}|i)))^2],$$

$$L_{i8} = \sum_{j \in \mathcal{C} \setminus \{i, c\}} \mathbb{E}_{G(\mathbf{z}|j) \sim p_j^g} [(1 - M_i(G(\mathbf{z}|j)))^2], \text{ and,}$$

$$L_{i9} = \mathbb{E}_{G(\mathbf{z}|i) \sim p_i^g} [(1 - D(G(\mathbf{z}|i)|i))^2].$$

3.5. Putting it all together

The model for the GAMO framework is detailed in Figure 3, while the complete algorithm is described in Algorithm 1. To ensure an unbiased training for M and D we generate artificial points for the i -th class with probability $(P_c - P_i)$ to compensate for the effect of imbalance. On the other hand, to also ensure unbiased training for G we use samples from all classes with equal probability.

4. Experiments

We evaluate the performance of a classifier in terms of two indices which are not biased toward any particular class [40], namely Average Class Specific Accuracy (ACSA) [21, 44] and Geometric Mean (GM) [26, 4]. All our experiments have been repeated 10 times to mitigate any bias generated due to randomization and the means and standard deviations of the index values are reported. Codes for the proposed methods are available at <https://github.com/SankhaSubhra/GAMO>.

We have used a collection of 7 image datasets for our experiments, namely MNIST [28], Fashion-MNIST [46], CIFAR10 [25], SVHN [34], LSUN [51] and SUN397 [47]. All the chosen datasets except SUN397 are not significantly imbalanced in nature, therefore we have created their imbalanced variants by randomly selecting a disparate number of samples from the different classes. Further, for all the datasets except SUN397, 100 points are selected from each class to form the test set. In the case of SUN397 (50 classes of which are used for our experiments) 20 points from each class are kept aside for testing.

We refrain from using pre-trained networks for our experiments as the pre-learned weights may not reflect the imbalance between the classes. We, instead, train the models

Table 1. Comparison of classification performance of CE and LS variants of classifiers on MNIST and Fashion-MNIST datasets.

Algorithm	MNIST				Fashion-MNIST			
	CE		LS		CE		LS	
	ACSA	GM	ACSA	GM	ACSA	GM	ACSA	GM
Baseline CN	0.88±0.01	0.87±0.02	0.88±0.01	0.86±0.01	0.82±0.01	0.80±0.01	0.81±0.01	0.79±0.01
SMOTE+CN	0.88±0.02	0.87±0.03	0.89±0.01	0.89±0.01	-	-	-	-
Augment+CN	-	-	-	-	0.82±0.01	0.78±0.01	0.82±0.01	0.78±0.01
DOS	-	-	-	-	0.82±0.01	0.79±0.01	0.81±0.01	0.79±0.02
(cGAN/cDCGAN)+CN	0.88±0.01	0.87±0.01	0.89±0.01	0.88±0.01	0.81±0.02	0.78±0.01	0.82±0.01	0.80±0.01
cG+CN	0.86±0.03	0.85±0.02	0.86±0.03	0.85±0.03	0.79±0.02	0.77±0.02	0.80±0.01	0.77±0.02
cG+D+CN	0.85±0.02	0.83±0.01	0.85±0.02	0.82±0.02	0.79±0.02	0.78±0.01	0.79±0.01	0.78±0.02
GAMO\D (Ours)	0.87±0.01	0.86±0.01	0.88±0.01	0.87±0.01	0.81±0.01	0.80±0.01	0.82±0.01	0.80±0.01
GAMO (Ours)	0.89±0.01	0.88±0.01	0.91±0.01	0.90±0.01	0.82±0.01	0.80±0.01	0.83±0.01	0.81±0.01

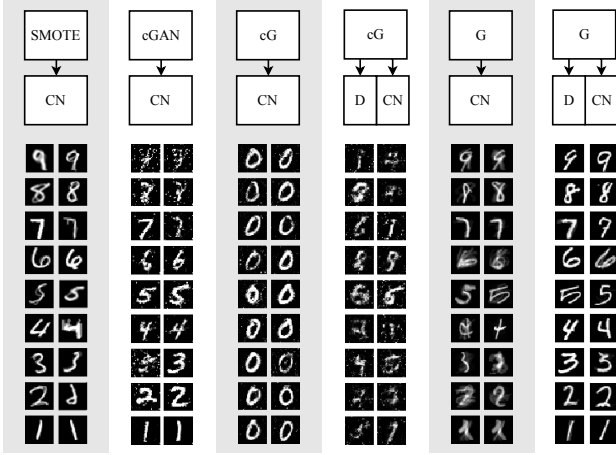


Figure 4. Ablation study on the MNIST dataset: SMOTE generates artificial samples from the minority class(es) as convex combinations of pairs of neighbors from the respective class(es). The oversampled dataset is then classified using a classifier network CN. SMOTE sometimes generates unrealistic “out-of-distribution” samples which are combinations of visually disparate images that happen to be Euclidean neighbors in the flattened image space. Using cGAN for generating new samples results in realistic images only from the more abundant minority classes. Training only a conditional Generator cG adversarially against CN, to generate images which will be misclassified by CN, results in new samples which all resemble the majority class ‘0’. Introducing a discriminator D (to ensure that cG adheres to class distributions) into the mix results in new samples which are somewhat in keeping with the class identities, but still unrealistic in appearance. Employing our proposed convex generator G to generate new samples by training it adversarially with CN (the GAMO\D formulation) results in samples which are in keeping with the class identities, but often “out-of-distribution” as the classes are non-convex. Finally, introducing D into this framework results in the complete GAMO model which can generate realistic samples which are also in keeping with the class identities.

from scratch to emulate real-world situations where the data is imbalanced and there is no pre-trained network available that can be used as an appropriate starting point. We have obtained the optimal architectures and hyperparameters for each contending method in Section 4-5 using a grid search (see supplementary document).

Table 2. Comparison of classification performance on CIFAR10 and SVHN datasets.

Dataset	Algorithm	ACSA	GM
CIFAR10	Baseline CN	0.45±0.01	0.37±0.01
	Augment+CN	0.47±0.01	0.39±0.02
	cDCGAN+CN	0.42±0.02	0.32±0.03
	DOS	0.46±0.02	0.37±0.01
	GAMO\D (Ours)	0.47±0.01	0.40±0.01
	GAMO (Ours)	0.49±0.01	0.43±0.02
SVHN	Baseline CN	0.74±0.01	0.73±0.01
	Augment+CN	0.69±0.01	0.63±0.01
	cDCGAN+CN	0.69±0.01	0.66±0.02
	DOS	0.71±0.02	0.68±0.01
	GAMO\D (Ours)	0.75±0.01	0.75±0.02
	GAMO (Ours)	0.76±0.01	0.75±0.02

4.1. MNIST and Fashion-MNIST

The experiments in this section are conducted using imbalanced subsets of the MNIST and Fashion-MNIST datasets. In case of both the datasets, we have sampled {4000, 2000, 1000, 750, 500, 350, 200, 100, 60, 40} points from classes in order of their index. Thus, the datasets have an Imbalance Ratio (IR: ratio of the number of representatives from the largest class to that of the smallest class) of 100. We begin by establishing the effectiveness of our proposed framework. We also compare between the two variants of GAMO which use Cross Entropy (CE) and Least Square (LS) losses, respectively.

We undertake an ablation study on MNIST using flattened images to facilitate straightforward visualization of the oversampled instances. Convolutional features are used for Fashion-MNIST. For MNIST, we have compared GAMO, against baseline classifier network (CN), SMOTE+CN (training set is oversampled by SMOTE), cGAN+CN (training set oversampled using cGAN, which is then used to train CN), and also traced the evolution of the philosophy behind GAMO, through cG+CN (conditional generator cG adversarially trained against CN, in contrast to cGAN+CN where CN does not play any part in training cGAN), cG+D+CN (cG+CN network coupled with a discriminator D), and GAMO\D (GAMO without a discriminator) on the MNIST dataset. SMOTE+CN and cGAN+CN are respectively replaced by Augment+CN (data augmenta-

Table 3. Comparison of classification performance with increased number of training instances on CelebA and LSUN datasets.

Algorithm	CelebA-Small				CelebA-Large			
	During Training		During Testing		During Training		During Testing	
	ACSA	GM	ACSA	GM	ACSA	GM	ACSA	GM
Baseline CN	0.91±0.01	0.91±0.01	0.59±0.01	0.45±0.04	0.93±0.01	0.92±0.01	0.71±0.01	0.60±0.03
Augment+CN	0.74±0.06	0.70±0.09	0.62±0.05	0.47±0.08	0.82±0.01	0.79±0.01	0.72±0.01	0.66±0.02
cDCGAN+CN	0.86±0.01	0.84±0.01	0.59±0.01	0.36±0.02	0.87±0.01	0.86±0.01	0.67±0.01	0.58±0.02
DOS	0.82±0.03	0.80±0.02	0.61±0.01	0.48±0.02	0.84±0.01	0.83±0.02	0.72±0.01	0.64±0.02
GAMO (Ours)	0.92±0.01	0.91±0.01	0.66±0.01	0.54±0.02	0.91±0.01	0.91±0.01	0.75±0.01	0.70±0.02

Algorithm	LSUN-Small				LSUN-Large			
	During Training		During Testing		During Training		During Testing	
	ACSA	GM	ACSA	GM	ACSA	GM	ACSA	GM
Baseline CN	0.90±0.01	0.89±0.01	0.50±0.01	0.28±0.05	0.87±0.01	0.87±0.01	0.61±0.02	0.54±0.03
Augment+CN	0.67±0.06	0.64±0.09	0.54±0.03	0.45±0.07	0.70±0.03	0.65±0.03	0.64±0.02	0.58±0.03
cDCGAN+CN	0.80±0.02	0.79±0.02	0.53±0.02	0.43±0.03	0.81±0.02	0.80±0.02	0.60±0.02	0.53±0.03
DOS	0.78±0.03	0.76±0.02	0.54±0.02	0.44±0.02	0.79±0.02	0.77±0.02	0.63±0.02	0.61±0.03
GAMO (Ours)	0.93±0.01	0.93±0.01	0.57±0.01	0.50±0.02	0.80±0.01	0.80±0.01	0.70±0.02	0.68±0.03

tion is used to create new images for balancing the training sets), and cDCGAN+CN (oversampled using conditional deep convolutional GAN) for Fashion-MNIST. GAMO is also compared with DOS on Fashion-MNIST.

The ablation study is shown visually in Figure 4 and the results for both datasets are tabulated in Table 1. Overall, GAMO is observed to perform better than all other methods on both datasets. Interestingly, GAMO\D performs much worse than GAMO on MNIST but improves significantly on Fashion-MNIST. This may be due to the fact that the convolutional feature extraction for Fashion-MNIST results in distributed representations where the classes are almost convex with little overlap between classes, enabling the convex generator to always generate data points which reside inside the class distributions.

Since we observe from Table 1 that the LS variants of the classifiers mostly perform better than their CE based counterparts (which according to [31] is contributed by the more stable and better decision boundary learned in LS), all the experiments in the subsequent sections are reported using the LS formulation for all the contending algorithms.

4.2. CIFAR10 and SVHN

In case of CIFAR10 and SVHN the classes are subsampled (4500, 2000, 1000, 800, 600, 500, 400, 250, 150, and 80 points are selected in order of the class labels) to achieve an IR of 56.25. From Table 2 we can see that GAMO performs better than others on both of these datasets, closely followed by GAMO\D, further confirming the additional advantage of convolutional feature extraction in the GAMO framework. Interestingly, Augment+CN performs much worse than the other methods on the SVHN dataset. This may be due to the nature of the images in the SVHN dataset, which may contain multiple digits. In such cases, attempting to augment the images may result in a shift of focus from one digit to its adjacent digit, giving rise to a discrepancy with the class labels.

Table 4. Comparison of classification performance on SUN397.

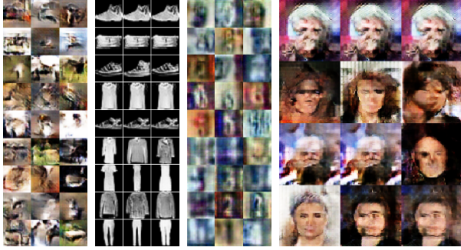
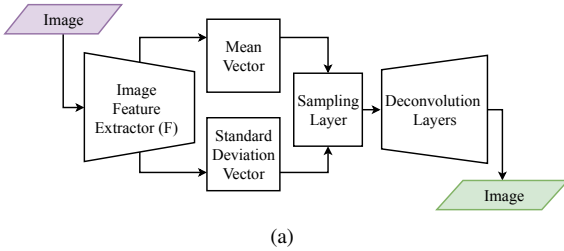
Algorithm	ACSA	GM
Baseline CN	0.26±0.04	0.19±0.05
Augment+CN	0.30±0.04	0.21±0.04
cDCGAN+CN	0.20±0.05	0.00±0.00
DOS	0.28±0.04	0.20±0.05
GAMO (Ours)	0.32±0.04	0.24±0.03

4.3. CelebA and LSUN

The experiment on CelebA and LSUN are undertaken to evaluate the performance of GAMO on images of higher resolution, as well as to assess the effects of an increase in the number of instances from the different classes. In case of CelebA the images are scaled to 64×64 size, while for LSUN the same is done on a central patch of resolution 224×224 extracted from each image. In the case of CelebA we have created two 5 class datasets by selecting samples from non-overlapping classes of hair colors, namely *blonde*, *black*, *bald*, *brown*, and *gray*. The first dataset is the smaller one (having 15000, 1500, 750, 300, and 150 points in the respective classes) with an IR of 100, while the second one is larger (having 28000, 4000, 3000, 1500, and 750 points in the respective classes) with an IR of 37.33. Similarly, in the case of LSUN we select 5 classes namely *classroom*, *church outdoor*, *conference room*, *dining room*, and *tower*, and two datasets are created. The smaller one (with 15000, 1500, 750, 300, and 150 points from the respective classes) has an IR of 100, while the larger one (with 50000, 5000, 3000, 1500, and 750 points) has an IR of 66.67.

In Table 3, we present the ACSA and GM over both the training and test set for the small and large variants of the two datasets. We can observe that all the algorithms manage to close the gap between their respective training and testing performances as the size of the dataset increases. Moreover, while Augment+CN seems to have the lowest tendency to overfit (smallest difference between training and testing performances), GAMO exhibits a greater ability to retain good

performance on the test dataset.



(b)



(c)

Figure 5. (a) GAMO2pix network. (b)-(c) Comparison of images respectively generated by cDCGAN, and GAMO2pix for (left to right) CIFAR10, Fashion-MNIST, SVHN, and CelebA-Small.

4.4. SUN397

We have randomly selected 50 classes from SUN397 to construct a dataset containing 64×64 sized images (depending on the image size either a 512×512 or a 224×224 center patch is extracted, which is then scaled to 64×64) with an IR of 14.21. The experiment on SUN397 is performed to evaluate the performance of GAMO over a large number of classes. A scrutiny of the result tabulated in Table 4 reveals that despite all four contending techniques being severely affected by the complexity of the classes and the scarcity of data samples from many of the classes, GAMO is able to retain overall better performance than its competitors.

5. GAMO2pix

GAMO results ultimately in a classifier trained to properly classify samples from all the classes. However, some application may require that actual samples be generated by oversampling to form an artificially balanced dataset. While GAMO directly generates images if flattened images

are used, it only generates vectors in the distributed representation space (mapped by the convolutional layers) for the convolutional variant. Therefore, we also propose the GAMO2pix mechanism to obtain images from the GAMO-generated vectors in the distributed representation space.

Table 5. Comparison of FID of cDCGAN and GAMO2pix.

Dataset	GAMO2pix (Ours)	cDCGAN
Fashion-MNIST	0.75 ± 0.03	5.57 ± 0.03
SVHN	0.17 ± 0.02	0.59 ± 0.04
CIFAR10	1.59 ± 0.03	2.96 ± 0.03
CelebA-Small	11.13 ± 0.04	15.12 ± 0.05

Our network for generating images (as illustrated in Figure 5(a)) from the GAMO-generated vectors is inspired by the Variational Autoencoder (VAE) [23, 37]. VAE, unlike regular autoencoders, is a generative model which attempts to map the encoder output to a standard normal distribution in the latent space, while the decoder is trained to map samples from the latent normal distribution to images. We follow the design of a standard VAE in GAMO2pix, only replacing the encoder part with the convolutional feature extractor F trained by GAMO. The GAMO2pix network is trained separately for each class while keeping the encoder part fixed. Such a setting should learn the inverse map from the D -dimensional feature space induced by F to the original image space and consequently be able to generate realistic images of the concerned class given GAMO-generated vectors for that class as input.

We present the images respectively generated by cDCGAN and GAMO2pix on CIFAR10, Fashion-MNIST, SVHN and CelebA-Small in Figures 5(b)-5(c). We can see that GAMO2pix can indeed generate more realistic and diverse images, compared to cDCGAN which also suffers from mode collapse for minority classes. This is further confirmed by the lower Fréchet Inception Distance (FID) [20] (calculated between real and artificial images from each class and averaged over classes) achieved by GAMO2pix, as shown in Table 5.

6. Conclusions and Future Work

The proposed GAMO is an effective end-to-end over-sampling technique for handling class imbalance in deep learning frameworks. Moreover, it is also an important step towards training robust discriminative models using adversarial learning. We have observed from our experiments that the convolutional variant of GAMO is more effective due to the distributed representations learned by the convolutional layers. We also found that the LS loss variant of GAMO generally performs better than the CE loss variant.

An interesting area of future investigation is to improve the quality of the images generated by GAMO2pix by employing a different architecture such as BEGAN [3]. To reduce the tendency of GAMO to overfit as well as to poten-

tially improve its performance, one may consider hybridization with improved GAN variants [16] which can achieve good performance even with less number of training samples. Further, one may explore the efficacy of GAMO to learn new classes by taking inspiration from Memory Replay GAN [45], or study the usefulness of the proposed convex generator for handling boundary distortion in GANs.

References

- [1] S. Ando and C. Y. Huang. Deep over-sampling framework for classifying imbalanced data. In *Machine Learning and Knowledge Discovery in Databases*, pages 770–785. Springer International Publishing, 2017. 2
- [2] S. Barua, M. M. Islam, X. Yao, and K. Murase. Mwmote—majority weighted minority oversampling technique for imbalanced data set learning. *IEEE Transactions on Knowledge and Data Engineering*, 26(2):405–425, 2014. 2
- [3] D. Berthelot, T. Schumm, and L. Metz. Began: boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017. 8
- [4] P. Branco, L. Torgo, and R. P. Ribeiro. A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys (CSUR)*, 49(2):31, 2016. 1, 5
- [5] M. Buda, A. Maki, and M. A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018. 1
- [6] S. R. Bulò, G. Neuhold, and P. Kotschieder. Loss max-pooling for semantic image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2126–2135, 2017. 1
- [7] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap. Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In *Advances in Knowledge Discovery and Data Mining*, pages 475–482, 2009. 2
- [8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002. 2
- [9] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer. Smoteboost: Improving prediction of the minority class in boosting. In *Knowledge Discovery in Databases: PKDD 2003*, pages 107–119, 2003. 2
- [10] Y.-A. Chung, H.-T. Lin, and S.-W. Yang. Cost-aware pre-training for multiclass cost-sensitive deep learning. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI’16*, pages 1411–1417. AAAI Press, 2016. 1
- [11] S. Das, S. Datta, and B. B. Chaudhuri. Handling data irregularities in classification: Foundations, trends, and future challenges. *Pattern Recognition*, 81:674–693, 2018. 1
- [12] Q. Dong, S. Gong, and X. Zhu. Imbalanced deep learning by minority class incremental rectification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 1
- [13] G. Douzas and F. Bacao. Effective data generation for imbalanced learning using conditional generative adversarial networks. *Expert Systems with applications*, 91:464–471, 2018. 2
- [14] A. Fernández, S. Garcia, F. Herrera, and N. V. Chawla. Smote for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research*, 61:863–905, 2018. 2
- [15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2
- [16] S. Gurumurthy, R. Kiran Sarvadevabhatla, and R. Venkatesh Babu. Deligan: Generative adversarial networks for diverse and limited data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 166–174, 2017. 9
- [17] H. Han, W.-Y. Wang, and B.-H. Mao. Borderline-smote: A new over-sampling method in imbalanced data sets learning. In *Advances in Intelligent Computing*, pages 878–887, 2005. 2
- [18] H. He, Y. Bai, E. A. Garcia, and S. Li. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *IEEE International Joint Conference on Neural Networks*, pages 1322–1328, 2008. 2
- [19] H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009. 1
- [20] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems 30*, pages 6626–6637, 2017. 8
- [21] C. Huang, Y. Li, C. Change Loy, and X. Tang. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5375–5384, 2016. 1, 5
- [22] S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*, 29(8):3573–3587, 2018. 1
- [23] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 8
- [24] B. Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016. 1
- [25] A. Krizhevsky. Learning multiple layers of features from tiny images. Technical Report TR-2009, University of Toronto, 2009. 5
- [26] M. Kubat, S. Matwin, et al. Addressing the curse of imbalanced training sets: one-sided selection. In *Icml*, volume 97, pages 179–186, 1997. 5
- [27] A. Kumar, P. Sattigeri, and T. Fletcher. Semi-supervised learning with gans: Manifold invariance with improved inference. In *Advances in Neural Information Processing Systems*, pages 5534–5544, 2017. 2
- [28] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 5

- [29] M. Lin, K. Tang, and X. Yao. Dynamic sampling approach to training neural networks for multiclass imbalance classification. *IEEE Transactions on Neural Networks and Learning Systems*, 24(4):647–660, 2013. 2
- [30] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollr. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007. IEEE, 2017. 1
- [31] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2017. 5, 7
- [32] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, and G. D. Tourassi. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural networks*, 21(2-3):427–436, 2008. 1
- [33] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2
- [34] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, page 5, 2011. 5
- [35] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier gans. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2642–2651. JMLR. org, 2017. 2
- [36] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 2
- [37] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014. 8
- [38] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016. 2
- [39] S. Santurkar, L. Schmidt, and A. Madry. A classification-based study of covariate shift in gan distributions. In *International Conference on Machine Learning*, pages 4487–4496, 2018. 2
- [40] M. Sokolova and G. Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427 – 437, 2009. 5
- [41] J. T. Springenberg. Unsupervised and semi-supervised learning with categorical generative adversarial networks. *arXiv preprint arXiv:1511.06390*, 2015. 2
- [42] A. Srivastava, L. Valkov, C. Russell, M. U. Gutmann, and C. Sutton. Veegan: Reducing mode collapse in gans using implicit variational learning. In *Advances in Neural Information Processing Systems 30*, pages 3308–3318, 2017. 2
- [43] S. Wang, W. Liu, J. Wu, L. Cao, Q. Meng, and P. Kennedy. Training deep neural networks on imbalanced data sets. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 4368–4374. IEEE, 2016. 1
- [44] Y.-X. Wang, D. Ramanan, and M. Hebert. Learning to model the tail. In *Advances in Neural Information Processing Systems*, pages 7029–7039, 2017. 2, 5
- [45] C. Wu, L. Herranz, X. Liu, J. van de Weijer, B. Raducanu, et al. Memory replay gans: Learning to generate new categories without forgetting. In *Advances in Neural Information Processing Systems*, pages 5966–5976, 2018. 9
- [46] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. 5
- [47] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, pages 3485–3492, 2010. 5
- [48] S. Xie and Z. Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015. 1
- [49] S. Xie and Z. Tu. Holistically-nested edge detection. *International Journal of Computer Vision*, 125(1-3):3–18, 2017. 1
- [50] Y. Yan, M. Chen, M. Shyu, and S. Chen. Deep learning for imbalanced multimedia data classification. In *2015 IEEE International Symposium on Multimedia (ISM)*, pages 483–488. IEEE, 2015. 1
- [51] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 5