Supervised Class Distribution Learning for GANs-based Imbalanced Classification

Zixin Cai[†], Xinyue Wang[†], Mingjie Zhou[‡], Jian Xu[†] and Liping Jing[†*]
[†]*Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing 100044, China*
Email: {18120340, 17112078, 17120432, lpjing}@bjtu.edu.cn
[‡]*Department of Mathematics, Hong Kong Baptist University, Hong Kong, China*
Email: 18481558@life.hkbu.edu.hk

***Abstract*— Class imbalance is a challenging problem in many real-world applications such as fraudulent transactions detection in finance and diagnosis of rare diseases in medicine, which has attracted more and more attention in the community of machine learning and data mining. The main issue is how to capture the fundamental characteristics of the imbalanced data distribution. In particular, whether the hidden pattern can be truly mined from minority class is still a largely unanswered question after all it contains limited instances. The existing methods provide only a partial understanding of this issue and result in the biased and inaccurate classifiers. To overcome this issue, we propose a novel imbalanced classification framework with two stages. The first stage aims to accurately determine the class distributions by a supervised class distribution learning method under the Wasserstein auto-encoder framework. The second stage makes use of the generative adversarial networks to simultaneously generate instances according to the learnt class distributions and mine the discriminative structure among classes to train the final classifier. This proposed framework focuses on Supervised Class Distribution Learning for Generative Adversarial Networks-based imbalanced classification (SCDL-GAN). By comparing with the state-of-the-art methods, the experimental results demonstrate that SCDL-GAN consistently benefits the imbalanced classification task in terms of several widely-used evaluation metrics on five benchmark datasets.**

***Keywords*-Imbalanced Classification; Class Distribution Learning; Generative Adversarial Networks**

I. Introduction

Class imbalance is an inevitable and challenging problem in various real-world applications such as telecommunication managements, bioinformatics, fraud detection, medical diagnosis and so on [1]. This problem occurs when the classes do not have equal number of training instances, especially when there is a big variance among the class sizes, which is usually caused by the rarity of events or by limitations on data collection process such as high cost or privacy problems. In this case, the traditional classification methods are always biased toward the majority class during training process, and affect the prediction performance on minority class. However, the minority class usually plays an important role on the final decision making. For example, in medical diagnosis, doctors usually prefer to high accuracy on cancer patients relative to healthy people [2]. Class imbalance has been considered as one of the top 10 problems in data mining [3], [4].

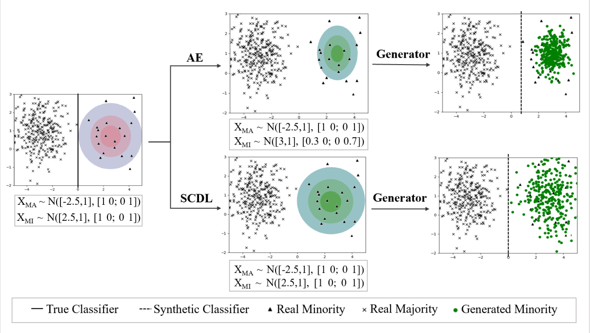


Figure 1. Illustration of a ”toy” two-class dataset with 300 majority-class instances (X_{MA}) and 20 minority-class instances (X_{MI}). The top part demonstrates the performance of autoencoders (AE) on capturing the class distribution. The below indicates the performance of our supervised class distribution learning method (SCDL).

There is an immense effort in the community of data mining and machine learning to develop imbalanced classification methods at data-level [5], [6] or algorithm-level [7]. Unfortunately, the limited data representation ability hinders these methods to achieve a comparably good result. Consequently, the deep neural network methods are presented to significantly boost imbalanced classification performance by adding crafted non-linear features into the cost-sensitive or sampling-based learning process [8]–[13]. However, these methods suffer from tuning the suitable cost variables or choosing the proper sampling seeds and neighborhoods [14]. Inspired by generative adversarial networks (GANs), researchers investigate the utility of using GANs to oversampling-based imbalanced classification [14]–[16]. Unfortunately, this attempt can lead to boundary distortion and further destroy the performance on the majority class, because it is hard for GANs to learn the true minority class distribution with limited instances [17].

[*] Corresponding author.

This work was supported in part by the National Natural Science Foundation of China under Grant 61822601, 61773050,and 61632004; the Beijing Natural Science Foundation under Grant Z180006; the Beijing Municipal Science & Technology Commission under Grant Z181100008918012; National Key Research and Development Program (2017YFC1703506); the Fundamental Research Funds for the Central Universities (2019JBZ110).

2374-8486/19/$31.00 ©2019 IEEE
DOI 10.1109/ICDM.2019.00014

Recently, Mariani et al. [17] tries to solve this problem via an autoencoder (AE)-based initialization strategy for GANs. Even though it has ability to infer the distributions of different classes in the latent space, AE can not capture the real class distribution since it minimizes only the reconstruction cost [18]. As shown in Fig. 1, given the imbalanced two-class dataset, the minority class distribution output by AE is different from the true distribution, esp., the mean $(3, 1)$ is far away from the true mean $(2.5, 1)$, and the covariance matrix is skewed (from [1 0; 0 1] to [0.3 0; 0 0.7]). It is clear that this learned distribution is not proper to augment the minority-class instances and further not properly capture the boundary regions among classes.

In this paper, therefore, we propose a supervised class distribution learning method for GANs-based imbalanced classification in a two-stage framework (SCDL-GAN). Specifically, in the first stage, a supervised class distribution learning model (SCDL) is designed under the Wasserstein auto-encoder architecture. As shown in Fig. 1, SCDL has ability to represent the distributive characteristics of all classes under the guideline of label information. In the second stage, the learned class distributions are taken as the initialization of GANs to start the adversarial training from a more stable point. Meanwhile, the discriminator is built on $K + 1$ classes to check if the instance is *fake* or belongs to one of the given K classes. The main contributions of this work can be summarized as follows.

- A new imbalanced class distribution learning model is proposed under the supervision of label information, which has ability to determine the intrinsic structure among different classes.
- An effective imbalanced classification framework is proposed by leveraging the strengths of Wasserstein auto-encoder and GANs, which has ability to simultaneously generate instances lying in the support of true distribution and identify the accurate boundary between classes.
- Extensive experiments on five widely-used benchmark datasets have shown that SCDL-GAN has ability to produce good performance on the minority class, while maintaining a reasonable overall accuracy.

The remainder of this paper is organized as follows. Section II lists the related works. Section III introduces the proposed SCDL-GAN framework. A series of experiments are conducted and discussed in Section IV. Section V briefly gives the conclusion and future work.

II. Related Work

The class imbalance problem has posed a significant drawback of the performance achieved by traditional classification system. In most real-world fields, this issue is particularly crucial since learning from these imbalanced data can help us discover useful knowledge to make important decisions while it can also be extremely costly to misclassify these data. A fundamental difficulty of this problem is the hardness of representing the distribution of the minority class because of domination of the majority class.

To solve this deficiency, in literatures, researchers have proposed to incorporate sampling and cost-sensitive into standard classification algorithms to improve the classification accuracy of the minority class [1]. Cost-sensitive methods assume that the misclassification of minority class instance is more expensive than the majority class instance, so that they adjust the learning process by assigning a higher cost value to the minority class [7]. A second option is to apply sampling methods aiming to balance data by undersampling on majority class or oversampling on minority class, which has shown great potential for imbalanced classification [5], [6]. To leverage the strength of deep learning, these two strategies are introduced to deep neural network models for imbalanced learning. For examples, Dong et al. [19] focuses on cost tuning to assign suitably higher costs to the minority instances. Another interesting kind of methods apply the oversampling to construct balanced dataset for deep learning [12], [13], [19]. However, these methods suffer from tuning the suitable parameters including cost value, sampling seeds and etc.

To emancipate the existing methods from tuning parameters, the powerful generative adverversarial networks (GANs) [20] is introduced in imbalanced classification. One typical method is adversarially re-weighting instances (ARIC) [21],which trains a network in an adversarial manner to automatically get the weights of majority class instances. ARIC obtains promising results on binary classification, but it can not directly handle multi-class task. The other is adopting GANs to generate sufficient minority class instances and build classifier on the augmented dataset [15], [22]. Once augmenting the original imbalanced data, they adopted the existing balanced classification methods to train the classifier. Such two-stage processing can not guarantee the generated data in the first stage is beneficial to training the subsequent classifier [14].

Recently, Mullick et al. [14] proposed an end-to-end oversampling deep imbalanced classification model, generative adversarial minority oversampling (GAMO). It effectively integrates the data generation and classifier training together with the aid of a three-player adversarial game including a convex generator, a classifier network, and a conditional discriminator. When generating minority class instances, GAMO uses a class-specific instance generation unit to reduce the computation complexity. Unfortunately, such strategy will destroy the performance of GAMO due to the lack of minority class information. To make up for the rare information of minority class, Mariani et al. [17] presented a balancing generative adversarial network (BAGAN) by exploiting all classes to estimate the class distributions in the latent space with the encoder module of an autoencoder (AE). However, it is hard to capture the complex structure

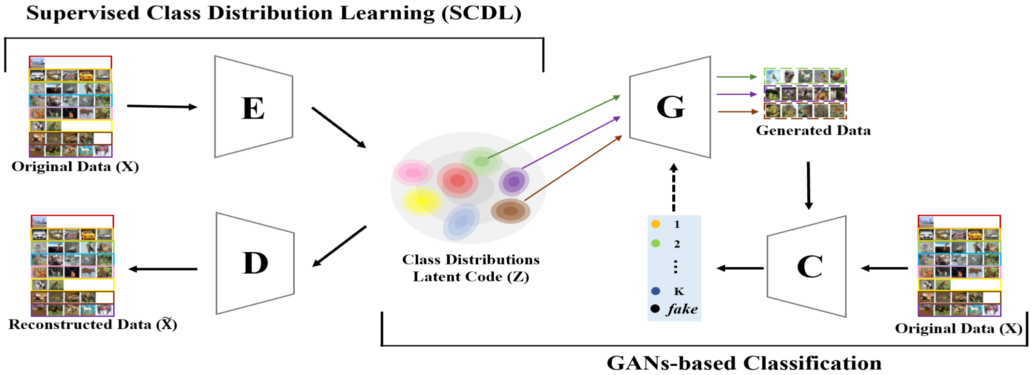


Figure 2. Framework of the proposed supervised class distribution learning for GANs-based imbalanced classification (SCDL-GAN).

from the limited data by only minimizing the reconstruction cost. Specifically, the instances may be encoded into non-overlapping zones chaotically scattered all across the latent space. In this case, the encoder module does not provide a useful representation and sampling from the latent space becomes hard [18]. As shown in Fig.1, the low-quality class distribution output by autoencoder will generate instances which do not fall into the support of true class distribution.

Therefore, in this paper, we propose a supervised class distribution learning method to accurately learn the class representations to initialize the subsequent GANs, which is expected to benefit the imbalanced classification.

III. Proposed Method

Given an imbalanced multi-class training dataset $\mathcal{D} = \{X, Y\} = \{(x_i, y_i)|x_i \in \mathcal{R}^m, y_i \in \{1, ..., K\}, i = 1, ..., n\}$, where X is an observed dataset in the input space $\mathcal{X}$, our goal is to build a deep model to learn the distributive characteristics of K classes and determine the clear boundaries between them.

To reach this goal, we propose a new imbalanced classification framework with two stages, as shown in Fig. 2, one for supervised class distribution learning (SCDL) and the other for GANs-based classification. In SCDL, an encoder module (E) encodes the input instances $X \in \mathcal{X}$ to latent codes $Z \in \mathcal{Z}$ ($\mathcal{Z}$ is the latent space), and a decoder (D) reconstructs the instances $\tilde{X}$ from Z. SCDL will estimate the class distributions in the latent space. In GANs, a generator (G) generates fake instances given Z , and a classifier (C) determines the margins between classes. The network structure of D and G are the same, and the network structure of E and C are almost the same except for the last layer. The last layer of E is a fully connected layer with an output of latent code. The last layer of C is a dense layer with a softmax activation on $K + 1$ neurons. GANs will output a multi-class classifier to predict the probability that the coming instance is *fake* or belongs to one of the real K *classes*. The whole framework SCDL-GAN aims to leverage the strengths of SCDL and GANs for imbalanced classificaiton. Next, we will describe SCDL-GAN in detail.

A. Supervised Class Distribution Learning (SCDL)

For high dimensional data, its main structures are often embedded in some low dimensional manifold [23], [24], which motivates us to design SCDL for capturing the class distributions from the latent low-dimensional space instead of the input high-dimensional space. Due to the lack of instances, it is hard to learn the minority class distribution. Inspired by BAGAN [17], we jointly utilize all given instances including majority and minority classes to exploit the available information as much as possible. With the aid of label information, SCDL tries to estimate the class distributions under Wasserstein auto-encoder framework (WAE) [18]. Here WAE is adopted because it allows non-random encoders to deterministically map inputs to their latent codes, which will benefit the class distribution estimation.

Specifially, we approach the generative modeling of SCDL from the optimal transport point of view under the auto-encoder framework. The parameters in model can be determined by minimizing the certain discrepancy measures between the true (but unknown) data distribution P_X and the reconstructed data distribution $P_{\tilde{X}}$ with the aid of Wasserstein distance. It has been proved that minimizing the Wasserstein distance is equal to optimizing the following regularized upper bound [18]:

$$\begin{gathered}\mathcal{L}_{WAE}(P_X, P_{\tilde{X}}) = \inf_{Q(Z|X)\in\mathcal{E}} \mathcal{L}_{recon} + \lambda_1 \cdot \mathcal{L}_{match} \\ \mathcal{L}_{recon} = \mathbb{E}_{P_X}\mathbb{E}_{Q(Z|X)}[c(X, \tilde{X})] \\ \mathcal{L}_{match} = \mathcal{DIV}_Z(Q_Z, P_Z)\end{gathered} \tag{1}$$

where $\mathcal{E}$ is any nonparametric set of the probabilistic encoders. λ_1 is a tradeoff between $\mathcal{L}_{recon}$ and $\mathcal{L}_{match}$. $c(p, q)$ is the cost function to measure the distance between p and q, here the squared cost function $c(p, q) = ||p - q||^2$ is used. $Q_Z = \mathbb{E}_{P_X}[Q(Z|X)]$ is the encoded distribution of training instances. P_Z is the prior distribution of latent codes Z. $\mathcal{DIV}_Z$ is an arbitrary divergence between Q_Z and P_Z. Here, Maximum Mean Discrepancy (MMD), a distance on the space of densities, is adopted since it shares the properties of divergence functions and has the ability to form an unbiased U-estimator [25]. The other reason is that it can be used in conjunction with stochastic gradient descent (SGD) methods.

The original WAE can be taken as an unsupervised learning process because it only uses X, which may result in overlapping or fuzzy margin among classes. Actually, the label information Y of training data is given, which is precious and helpful to determine the intrinsic structures of classes [26]. Thus, we design the following cross-entropy regularizer to exploit label information,

$$\mathcal{L}_{SV} = -\mathbb{E}_{E(Z|X)}[Y_X^T \log(P(Y_X|\tilde{X}))] \tag{2}$$

here Y_X indicates the label of the given training instance X. This term aims to enforce the reconstructed data $\tilde{X}$ having the same label information with the original data X.

By combining (1) and (2), we can get the final objective of SCDL as follows.

$$\mathcal{L}_{SCDL} = \mathcal{L}_{WAE}(P_X, P_{\tilde{X}}) + \lambda_2 \cdot \mathcal{L}_{SV} \tag{3}$$

where λ_2, similar to λ_1, is a tradeoff between $\mathcal{L}_{SV}$ and $\mathcal{L}_{WAE}(P_X, P_{\tilde{X}})$. This model can be optimized with mini-batch ADAM [27]. The proposed SCDL can simultaneously keep the good properties of WAE (such as stable training, encoder-decoder architecture, nice latent manifold structure) and identify more discriminative characteristics among training data. Suppose the k-th class follows a multivariate normal distribution $Z_k = \mathcal{N}(\mu_k, \sigma_k)$ in the latent space with mean vector μ_k and covariance matrix σ_k, the original instances of the k-th class will be fed into the encoder module (E) to get their latent representations, i.e., $Z_k = E(X_k)$. As shown in Fig. 1 (the middle part), SCDL is able to accurately estimate the class distributions for both majority class and minority class.

B. Adversarial Training for Imbalanced Classification

To generate realistic instances for all classes and identify the boundaries among different classes, SCDL-GAN is proposed to train a generator G and a classifier C in an adversarial manner. G aims to generate instances for K classes and C is required to label the instances either as *fake* or with one real class label from $\{1, \cdots, K\}$. In other words, C and G play a two-player minimax game with the following functions:

$$\mathcal{L}_G = -\mathbb{E}_{\tilde{x} \sim P_G(z)} \log\left[C(\tilde{x}, y)\right] \tag{4}$$

$$\mathcal{L}_C = -\mathbb{E}_{x \sim P_X(x)} \log\left[C(x, y)\right] - \mathbb{E}_{\tilde{x} \sim P_G(z)} \log\left[C(\tilde{x}, \tilde{y})\right] \tag{5}$$

where $y \in \{1, \cdots, K\}$ indicates the real class label and $\tilde{y} = K + 1$ is for the *fake* class. During the adversarial training, G and C are fine tuned to minimize their corresponding loss functions. To learn differences between fake instances and real instances from different classes, C is trained with real and fake instances jointly.

Similar to SCDL, the above GANs model is optimized by the mini-batch ADAM. To avoid the "mode collapse" problem and make GANs start from a stable point, we transfer the knowledge learnt by SCDL to GANs. Specifically, the generator G is initialized with the weights in the decoder module (D) and the first layers of classifier C with the weights of the encoder module (E). The inputs of G are latent codes randomly sampled from the class distribution of target class learnt by SCDL. Meanwhile, during learning process, the class distributions $\{Z_k\}_{k=1}^K$ are considered invariant, which benefits forcing the generator (G) not to diverge from the learnt class encoding in the latent space. As shown in Fig. 1 (the right part), the generated instances output by G are consistent with the true class distribution.

To avoid imbalance problem, in each mini-batch when training C, $1/(1+K)$ of instances are fake and generated by G, and the rest instances are randomly sampled from K real classes. Fake instances are uniformly distributed among K classes, which further provides the best possible balance for the fake data.

During the adversarial learning, the classifier C will be trained to discriminate instances from fake and different real classes, while the parameters in G are optimized to generate fake instances for confusing C. Consequently, competition in this adversarial game drives both G and C to improve their performance until C learns clear margins between different real classes, but can't distinguish between real and fake instances. Once the classifier C is built, we can predict the label of new coming instance $\hat{x}$ by selecting the label with the largest value in $\{\hat{o}_k\}_{k=1}^{K+1} = C(\hat{x})$, i.e., $j = argmax_k \{\hat{o}_k\}_{k=1}^{K+1}$, where o_k is the value of k-th neuro in the output layer.

IV. Experiments

In this section, a series of experiments are conducted on five real-world benchmark datasets. The proposed imbalanced classification method (SCDL-GAN) is evaluated by comparing with the state-of-the-art methods.

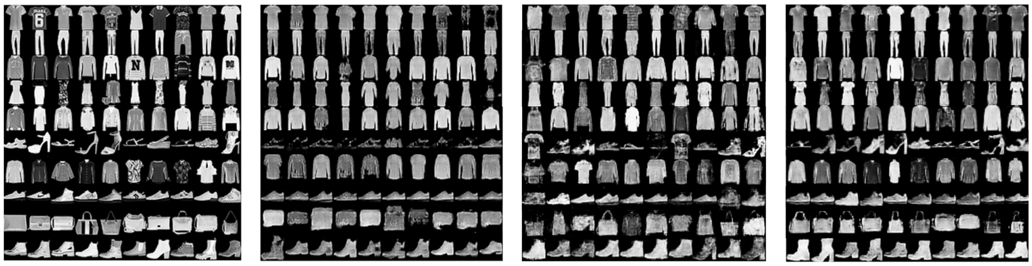
A. Experimental Settings

This subsection gives a brief introduction of experimental setting including benchmark datasets, evaluation metrics of imbalanced classification and the selected baselines with parameter settings.

Datasets: In experiments, five benchmark datasets, MNIST [28], Fashion-MNIST [29], CIFAR-10 [30], SVHN [31] and CelebA [32], are used to validate the imbalanced classification performance. The first two datasets are single-channel images, and the last three datasets are three-channel images which are relatively hard to analyze. For CelebA, the original images are resized to 64×64 due to reducing the computation complexity. Five non-overlapping classes along the target feature, hair color, are used including *blonde*,

Table I
SUMMARIZATION OF DATASETS IN EXPERIMENTS

Dataset name	**Shape**	**Classes**	**IR**	**Training Set** (♯Instances per Class)	**Testing Set** (♯Instances per Class)
MNIST	$28 \times 28 \times 1$	10	100	{4000,2000,1000,750, 500,350,200,100,60,40}	{980,1135,1032,1010, 982,892,958,1028,974,1009}
Fashion-MNIST	$28 \times 28 \times 1$	10	100	{4000,2000,1000,750, 500,350,200,100,60,40}	{1000,1000,1000,1000,1000, 1000,1000,1000,1000,1000}
CIFAR-10	$32 \times 32 \times 3$	10	56.25	{4500,2000,1000,800, 600,500,400,250,150,80}	{1014,1012,1023,1012,991, 1016,1005,1015,965,947}
SVHN	$32 \times 32 \times 3$	10	56.25	{4500,2000,1000,800, 600,500,400,250,150,80}	{1744,5099,4149,2882,2523, 2384,1977,2019,1660,1595}
CelebA	$64 \times 64 \times 3$	5	100	{15000,1500,750,300,150}	{2660,5422,412,3428,535}



(a) Real Image Instance (b) BAGAN (c) WAE-GAN (d) SCDL-GAN

Figure 3. The example real images and synthetic images: (a) real images from Fashion-MNIST and images generated by (b) BAGAN, (c) WAE-GAN and (d) the proposed SCDL-GAN on the imbalanced Fashion-MNIST dataset as shown in Table I.

black, *bald*, *brown* and *gray*. All selected datasets are not significantly imbalanced in nature, following [14], we create imbalanced variants by randomly selecting instances with different sizes from different classes in order of their indices. Details of these datasets are shown in Table I. Obviously, there is a large imbalance ratio (IR: ratio of the number of representatives from the largest class to that of the smallest class) in each dataset from 56.25 to 100. For each dataset, there is a public testing set which is directly used to evaluate the classification performance.

Evaluation Metrics: To account for all classes no matter majority or minority class, four evaluation metrics are adopted to validate the imbalanced classification performance, which are commonly used in imbalanced learning [11], [14], [33]. They are Average Class Specific Accuracy (ACSA), macro-averaged precision (PM), macro-averaged F-measure (FM) and macro-averaged geometric-mean (GM). Sokolova and Lapalme have proved that they are not biased toward any particular class [34]. Larger value of ACSA, PM, FM or GM indicates better performance. Due to the macro-average strategy is adopted, thus, these metrics treat all classes fairly and evaluate the classification performance on overall classes.

Baselines: Our goal is to improve the imbalanced classification performance by augmenting data, thus, the following three methods recently proposed are taken as baselines:

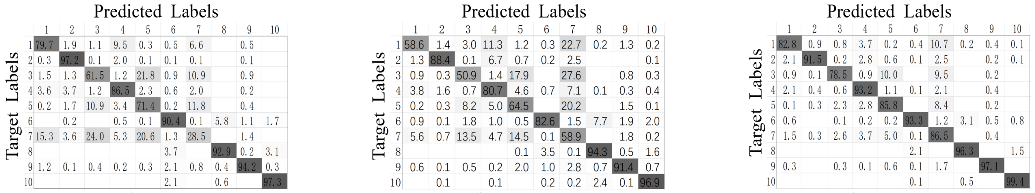
DFBS [13] is an oversampling-based deep learning method, where a two-layer CNN (cross entropy and triplet loss) is used to learn the discriminative feature space in which synthetic minority-class instances are generated to augment the original dataset. Then, *logistic regression* is used to train the classifier on the augmented dataset. In this case, DFBS can be taken as a two-stage method. It has been proven that DFBS outperforms the traditional oversampling methods, thus we select it as baseline rather than the existing SMOTE-type methods [5], [6].

BAGAN [17] takes advantage of generative adversarial network and autoencoder to augment the dataset. It first identifies the class distribution in latent space via autoencoder and then initializes GANs by the trained encoder and decoder. Subsequently, *RESNET-18* model [35] is adopted to train the multi-class classifier with augmented data. The experimental results show that BAGAN is better than the state-of-the-art ACGAN [36] on imbalanced data augmentation, thus we select it as baseline rather than ACGAN.

GAMO [14] is a one-stage imbalance classification method by integrating generator, multi-class classifier and fake/real discriminator together in a three-player adversarial game. Similar to DFBS, GAMO oversamples new instances in the latent space rather than the input feature space. GAMO empirically demonstrates its good performance by comparing with the two-stage methods including cDCGAN+CN, SMOTE+CN and the recent one-stage method DOS [12]. Thus, GAMO is taken as one baseline rather than DOS.

DFBS and BAGAN are two-stage methods, which only generate data and then use existing classification algorithms to train the final classifier. GAMO, like the proposed SCDL-GAN, trains the generator and classifier together in one framework.

Parameter Settings: The optimal experimental settings



Predicted Labels

Target Labels	1	2	3	4	5	6	7	8	9	10
1	79.7	1.9	1.1	9.5	0.3	0.5	6.6		0.5	
2	0.3	97.2	0.1	2.0	0.1	0.1	0.1		0.1	
3	1.5	1.3	61.5	1.2	21.8	0.9	10.9		0.9	
4	3.6	3.7	1.2	86.5	2.3	0.6	2.0		0.2	
5	0.2	1.7	10.9	3.4	71.4	0.2	11.8		0.4	
6		0.2		0.5	0.1	90.4	0.1	5.8	1.1	1.7
7	15.3	3.6	24.0	5.3	20.6	1.3	28.5		1.4	
8						3.7		92.9	0.2	3.1
9	1.2	0.1	0.4	0.2	0.3	2.1	0.8	0.4	94.2	0.3
10						2.1		0.6		97.3

(a) BAGAN

Predicted Labels

Target Labels	1	2	3	4	5	6	7	8	9	10
1	58.6	1.4	3.0	11.3	1.2	0.3	22.7	0.2	1.3	0.2
2	1.3	88.4	0.1	6.7	0.7	0.2	2.5			0.1
3	0.9	0.3	50.9	1.4	17.9		27.6		0.8	0.3
4	3.8	1.6	0.7	80.7	4.6	0.7	7.1	0.1	0.3	0.4
5	0.2	0.3	8.2	5.0	64.5		20.2		1.5	0.1
6	0.9	0.1	1.8	1.0	0.5	82.6	1.5	7.7	1.9	2.0
7	5.6	0.7	13.5	4.7	14.5	0.1	58.9		1.8	0.2
8					0.1	3.5	0.1	94.3	0.5	1.6
9	0.6	0.1	0.5	0.2	2.0	1.0	2.8	0.7	91.4	0.7
10		0.1		0.1		0.2	0.2	2.4	0.1	96.9

(b) WAE-GAN

Predicted Labels

Target Labels	1	2	3	4	5	6	7	8	9	10
1	82.8	0.9	0.8	3.7	0.2	0.4	10.7	0.2	0.4	0.1
2	2.1	91.5	0.2	2.8	0.6	0.1	2.5		0.2	0.1
3	0.9	0.1	78.5	0.9	10.0		9.5		0.2	
4	2.1	0.4	0.6	93.2	1.1	0.1	2.1		0.4	
5	0.1	0.3	2.3	2.8	85.8		8.4		0.2	
6	0.6		0.1	0.2	0.2	93.3	1.2	3.1	0.5	0.8
7	1.5	0.3	2.6	3.7	5.0	0.1	86.5		0.4	
8						2.1		96.3		1.5
9	0.3		0.3	0.1	0.6	0.1	1.7		97.1	
10						0.1		0.5		99.4

(c) SCDL-GAN

Figure 4. Mode collapse analysis for BAGAN, WAE-GAN and the proposed SCDL-GAN on Fashion-MNIST dataset.

for each method are determined either by experiments or suggested by previous works. In supervised class distribution learning, we adaptively adjust the parameters λ_1 and λ_2 in each epoch:

$$\lambda_1^{t+1} = 0.5 \times \lambda_1^t + 0.25 \times (\frac{\mathcal{L}_{match}^t}{\mathcal{L}_{recon}^t} + \frac{\mathcal{L}_{match}^t}{\mathcal{L}_{SV}^t}) \tag{6}$$

$$\lambda_2^{t+1} = 0.5 \times \lambda_2^t + 0.25 \times (\frac{\mathcal{L}_{SV}^t}{\mathcal{L}_{recon}^t} + \frac{\mathcal{L}_{SV}^t}{\mathcal{L}_{match}^t}) \tag{7}$$

where the superscript t is the index of epoch and $\lambda_1^1 = \lambda_2^1 = 1$. Their values in each epoch depend on the corresponding loss.

B. Generated Instances Assessment

In this experiment, our goal is to check the quality of generated instances. DFBS and GAMO generate artificial instances in the latent space and do not convert them back to the original feature space, thus, it is hard to directly compare the proposed method with them. Besides, in order to check the effect of cross-entropy regularizer $\mathcal{L}_{SV}$, we also check the qulity of instances generated by WAE-GAN, which is the same as SCDL-GAN except it without cross-entropy regularizer. Fig. 3 gives the examples of generated instances by applying BAGAN, WAE-GAN and the proposed SCDL-GAN on Fashion-MNIST dataset.

As expected, SCDL-GAN can indeed generate more realistic and diverse images for each class than BAGAN and WAE-GAN. Take the eighth class about *sneaker* as an example, the artificial instances output by BAGAN look similar, while SCDL-GAN is able to generate various *sneaker*. As for WAE-GAN, in the 6-th class about *Sandal* and 9-th class about *Bag*, there are generated instances look similar to *T-shirts*.

Among ten classes, the 3-rd class about *Pullover*, 5-th class about *Coat* and 7-th class about *Shirt* are relatively similar and hard to distinguish, i.e., the class distribution is not easy to be characterized. From Fig. 3(b) and (c), it can be seen that the artificial instances of these three classes (output by BAGAN and WAE-GAN) are close to each other, which will result in fuzzy boundary between classes and make classification in a more difficult situation. Fortunately, as shown in Fig. 3(c), there are significant differences among these three classes, where the instances are generated by the proposed SCDL-GAN. Moreover, with the decreasing of class size (from the 1-st class about T-shirt/top to the 10-th class about Ankle boot), the quality of generated images by SCDL-GAN keeps satisfactory, however BAGAN and WAE-GAN output worse and worse artificial images. This result confirms the ability of SCDL-GAN on generating high-quality instances for imbalanced classification.

C. Mode Collapse Analysis

The above results about generated images motivate us to investigate the generative adversarial training process in BAGAN, WAE-GAN and SCDL-GAN. It can be seen that BAGAN and WAE-GAN does not obtain good performance on some classes. We believe the main reason is that BAGAN suffers from mode collapse and boundary distortion due to the limitation of the initialization strategy (autoencoder and WAE). To confirm this, following [37], we apply covariate shift analysis (CSA) on the generators of BAGAN, WAE-GAN and SCDL-GAN. For imbalanced classification, we assume that "If GAN is trained on an imbalanced dataset and it can reproduce this discriminative structure, then it will not suffer from mode collapse". This analysis firstly trains a multi-class classifier on a balanced dataset $\mathcal{BD}$ and a GAN on an imbalanced dataset $\mathcal{UD}$. Secondly it generates a synthetic dataset with the same number of dataset $\mathcal{BD}$ by sampling images from the trained GANs. Then it uses the built classifier to predict labels for the synthetic dataset. Finally, we can obtain the quantitative assessment of the mode distribution by checking the matching score between predicted labels and ground-truth labels.

In experiments, the whole original Fashion-MNIST dataset is taken as $\mathcal{BD}$ including 60000 images evenly belonging to ten classes. Our experimental imbalanced dataset (as shown in Table I) is taken as $\mathcal{UD}$. Following BAGAN, *RESNET-18* is adopted as the multi-class classifier (C_{RESNET}) and trained on $\mathcal{BD}$ to accurately detect the hidden class structure. BAGAN, WAE-GAN and SCDL-GAN are all trained on the imbalanced dataset $\mathcal{UD}$ to train the generator (G) which is used to generate the artificial datasets (denoted as AD_{BAGAN} from BAGAN, $AD_{WAE-GAN}$ from WAE-GAN and $AD_{SCDL-GAN}$ from SCDL-GAN respectively) with the same number of $\mathcal{BD}$. Finally, the predicted labels on AD_{BAGAN}, $AD_{WAE-GAN}$ and $AD_{SCDL-GAN}$ are output by the classifier C_{RESNET}.

Figure 4 gives the confusion matrices of the prediction results on AD_{BAGAN}, $AD_{WAE-GAN}$ and $AD_{SCDL-GAN}$.

Table II
COMPARISON OF CLASSIFICATION PERFORMANCE ON SINGLE-CHANNEL IMAGESETS MNIST AND FASHION-MNIST.

	MNIST				**Fashion-MNIST**			
Method	***ACSA***	***PM***	***FM***	***GM***	***ACSA***	***PM***	***FM***	***GM***
DFBS [13]	0.85	0.89	0.85	0.89	0.77	0.84	0.77	0.85
GAMO [14]	0.9	0.89	0.89	0.91	0.83	0.83	0.81	0.89
BAGAN [17]	0.98	0.98	0.98	0.98	0.83	0.85	0.82	0.89
SCDL-GAN	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.93**	**0.91**	**0.95**

Table III
COMPARISON OF CLASSIFICATION PERFORMANCE ON THREE-CHANNEL IMAGESETS CIFAR-10, SVHN AND CELEBA.

	CIFAR-10				**SVHN**				**CelebA**			
Method	***ACSA***	***PM***	***FM***	***GM***	***ACSA***	***PM***	***FM***	***GM***	***ACSA***	***PM***	***FM***	***GM***
DFBS [13]	0.45	0.62	0.44	0.62	0.69	0.82	0.71	0.81	0.38	0.46	0.28	0.41
GAMO [14]	0.49	0.53	0.47	0.43	0.76	0.79	0.76	0.86	0.63	0.64	0.57	0.73
BAGAN [17]	0.59	0.67	0.58	0.74	0.83	0.85	0.83	0.90	0.72	0.76	0.70	0.77
SCDL-GAN	**0.99**	**0.97**	**0.95**	**0.98**	**0.99**	**0.96**	**0.95**	**0.98**	**0.98**	**0.98**	**0.96**	**0.97**

Row indicates the ground-truth label, while the column indicates the predicted label. For convenient demonstration, we normalize each row so that the summation of its cells is 100. The higher the value, the darker the corresponding grid in the confusion matrix. Obviously, the diagonal blocks in matrix (c) (for SCDL-GAN) are much darker than other matrics (for BAGAN and WAE-GAN), and much more instances fall into the non-diagonal cells in left matrix. Especially, the 3-rd class about *Pullover*, 5-th class about *Coat* and 7-th class about *Shirt* are definitely confused according to the artificial instances generated by BAGAN and WAE-GAN. Only 28.5% and 58.9% instances generated by BAGAN and WAE-GAN are correctly assigned to the 7-th class. As expected, SCDL-GAN has ability to generate class-separable artificial instances, thus, the confusion matrix looks much clearer than BAGAN and WAE-GAN. Similar results can be obtained on other four datasets. Due to page limitation, we do not list all results here.

This experimental result confirms that SCDL-GAN will not fall into the "mode collapse" problem. It benefits from that SCDL has ability to detect the discriminative class distribution in the latent space even though there are few instances in the minority classes.

D. Classification Performance Comparison

In this subsection, we compare SCDL-GAN with three state-of-the-art methods in terms of four evaluation metrics (ACSA, PM, FM and GM). Table II and Table III list the classification performance on two single-channel datasets and three three-channel datasets respectively. To be exciting, SCDL-GAN consistently achieves the best results in all evaluation metrics on all datasets. From this result, we can get the following observations:

DBFS obtains the worst performance on five datasets because it is a two-stage learning process, i.e., data generation is independent on the classifier building. This strategy can not guarantee the generated instances are helpful to create the margins among classes.

GAMO performs over DBFS, which benefits from its one-stage framework by integrating oversampling, discriminator and classifier together. However, it is hard to trade off the discriminator and classifier. As a result, the generator may be rewarded for generating images that look real but can not demonstrate the intrinsic structure of the corresponding class.

The generator in DBFS and GAMO is class-specific, i.e., the samples are generated only according to the current class. This will deteriorate the quality of artificial instances because the minority class can not provide sufficient information to learn the realistic data distribution.

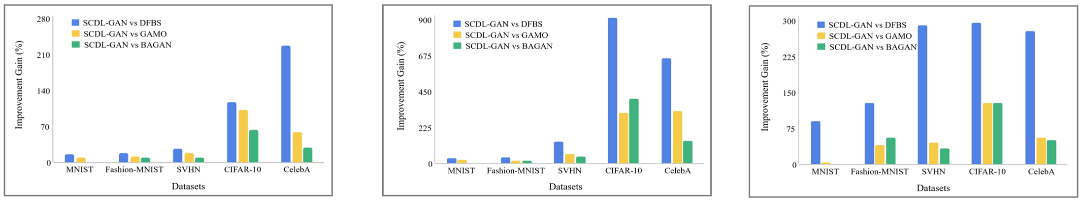
BAGAN is superior to the previous two methods because it takes into account all classes when determing the class distribution, which makes up for the lack of data in minority class to some extent. Unfortunately, BAGAN adopts the traditional autoencoder to initialize the generator and discriminator of the subsequent GAN model, which will push BAGAN fall into the "mode collapse" problem as shown in the last subsection.

As expected, the propsed SCDL-GAN outperforms all baselines on all experimental datasets. One reason is that the new supervised class distribution learning method is helpful to determine the intrinsic struture of data and discriminative information for each class, which benefits the subsequent GAN to generate realistic and diverse artificial instances. The other reason is that the generator and classifier are iteratively trained so that the generator correctly augments imbalanced data and the classifier effectively identifies the clear boundaries among classes.

Meanwhile, in real-world applications, the minority class is of more interest. Taking disease surveillance as an example, there are much fewer sick persons than healthy persons. Predicting a sick person to be healthy is always much more

Table IV
COMPARISON OF CLASSIFICATION PERFORMANCE IN TERMS OF RECALL ON THE SMALLEST CLASS (***REC***$_{MI}$) AND PRECISION ON THE LARGEST CLASS (***PRE***$_{MA}$) FOR EACH DATASET.

Method	MNIST		Fashion-MNIST		CIFAR-10		SVHN		CelebA	
	REC$_{MI}$	***PRE***$_{MA}$	***REC***$_{MI}$	***PRE***$_{MA}$	***REC***$_{MI}$	***PRE***$_{MA}$	***REC***$_{MI}$	***PRE***$_{MA}$	***REC***$_{MI}$	***PRE***$_{MA}$
DFBS [13]	0.74	0.5	0.71	0.42	0.1	0.25	0.41	0.24	0.12	0.25
GAMO [14]	0.78	0.9	0.83	0.68	0.24	0.43	0.6	0.63	0.21	0.6
BAGAN [17]	0.93	0.97	0.84	0.61	0.2	0.43	0.65	0.69	0.37	0.62
SCDL-GAN	**0.99**	**0.94**	**0.99**	**0.95**	**0.99**	**0.97**	**0.96**	**0.92**	**0.89**	**0.93**



(a) Macro-averaged F-measure (FM) on overall classes

(b) Recall on the smallest class (REC$_{\mathrm{MI}}$)

(c) Precision on the largest class (PRE$_{\mathrm{MA}}$)

Figure 5. The improvement gain of the proposed SCDL-GAN over three baselines on five benchmark datasets.

costly than classifying a healthy person as a patient. In other words, it is expected that the classifier can return a higher recall on the minority class and higher precision on the majority class. Thus, we also calculate recall on the smallest class (with minimum number of instances) and precision on the largest class (with maximum number of instances) in five datasets for four methods. Table IV gives the classification results obtained by SCDL-GAN and three baselines. It can be seen that, for Fashion-MNIST dataset, three baselines increase the performance on minority class but their performances on majority class are not signifcanlty improved. This confirms that the data generated by DFBS/GAMO/BAGAN may be confused between classes, as shown in Fig. 3. For CIFAR-10 and CelebA, baselines definitely output worse results on minority class, which indicates that they can not determine the clear boundary for minority class. Fortunately, nomatter minority class and majority class, SCDL-GAN consistently obtains the best and high-quality classification results.

To further demonstrate the superiority of SCDL-GAN, we calculate its improvement gain (IG) over three baselines for each evaluation matric via

$$IG = (R_{SCDL-GAN} - R_{Baseline})/R_{Baseline} \tag{8}$$

Here, $R_{SCDL-GAN}$ indicates the performance obtained by SCDL-GAN, and $R_{Baseline}$ represents the performance of the corresponding baseline.

In this experiments, we select the macro-averaged F-meansure to check the improvement gain on overall classes, Recall on the smallest class and Precision on the largest class. As shown in Figure 5(a) , on average, SCDL-GAN achieves an FM improvement of 17.33%, 11.79%, 6.00% over DFBS, GAMO and BAGAN respectively on simple single-channel imagesets, while obtains 123.03%, 60.50%, 34.48% over three baselines on three complex three-channel image sets. Similarly, SCDL-GAN significantly improves the classification performance on both minority and majority classes, as shown in Figure 5(b)-(c). Especially, it is hard to mine the hiden structure for CIFAR-10 dataset because it contains tiny images about natural scene. To be exciting, SCDL-GAN obtains more than 300% improvement gain over baselines on recall of the smallest class, and 125% on precision of the largest class. This demonstrates that SCDL-GAN not only significantly improves the classification performance of minority classes, but also effectively improves the classification performance of majority classes.

E. tSNE analysis

In order to investigate the learning process of each method, we adopt tSNE analysis [38] to visually present the discriminative ability. Taking datasets Fashion-MNIST and CIFAR-10 as examples, tSNE is firstly applied on the testing data, i.e. projecting the original feature space to a 2-dimension space. As shown in Fig. 6, it is not easy to directly separate the input data.

In experiments, DFBS uses the logistic regression as the classification method which will output a K-dimensional vector for each testing instance (K is the number of classes). GAMO and BAGAN adopt the neural networks with K neurons in the output layer, similar to DFBS, they obtain a K-dimensional vector for each testing instance. In SCDL-GAN, there are $K+1$ neurons in the output layer. Because the last neuron is used to check the probability that the instance is fake, only the first K neurons are used to represent the input instance for tSNE analysis. This vector can be taken as the final representation of testing data, thus, we use tSNE to project it on a 2-dimension space. When visualizing the instances, different classes are marked in

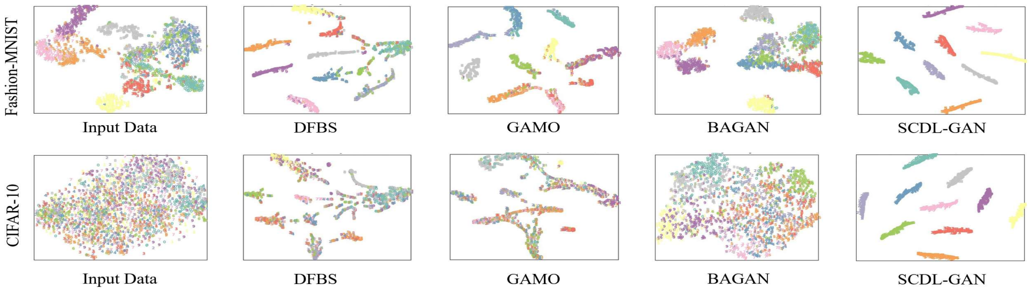


Figure 6. tSNE analysis on two datasets: Fashion-MNIST and CIFAR-10 dataset (Instances from different classes are marked by different colors).

different colors. The results are demonstrated in Fig. 6. As expected, SCDL-GAN outputs a compact representation for each class and effectively separates different classes. This further confirms that the proposed method has ability to extract the discriminative information from imbalanced data.

V. Conclusions and Future Work

In this paper, we present a supervised class distribution learning method for Generative Adversarial Networks (GANs)-based imbalanced classification, denoted as SCDL-GAN. Among it, SCDL is able to accurately capture the class distributions with the aid of supervised Wasserstein auto-encoder. The trained encoder module and decoder module can provide good initializations to the classifier and generator respectively in the subsequent generative adversarial networks. The adversarial training will enforce the generator to create high-quality and diverse artificial instances, meanwhile, guarantee the classifier to determine the discriminative characteristics for overall classes. A series of experiments have shown the superiority of SCDL-GAN on five benchmark datasets by comparing with the state-of-the-art baselines from several views.

In the current framework, the class distributions in latent space are kept invariant to force the generator consistent with the latent class distributions. This strategy works well by assuming that the learnt distribution approches to the true distribution. It will be interesting to update the class distributions in latent space during adversarial training so that they are close to the input distribution and benefits building the classifier. This work focuses on multi-class datasets where the classes are assumed non-overlapping. Actually, there are many multi-label imbalanced data, especially for extreme multi-label dataset where most instances fall into few classes but few instances belong to most classes. In this case, there may be complicated relationships among classes, how to extend the proposed method for such complex problem will be an interestiong topic.

References

[1] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Systems with Applications*, vol. 73, pp. 220–239, 2017.

[2] 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*, vol. 21, no. 2-3, pp. 427–436, 2008.

[3] Q. Yang and X. Wu, "10 challenging problems in data mining research," *International Journal of Information Technology & Decision Making*, vol. 5, no. 04, pp. 597–604, 2006.

[4] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221–232, 2016.

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[6] H. He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. IEEE, 2008, pp. 1322–1328.

[7] P. Branco, L. Torgo, and R. P. Ribeiro, "A survey of predictive modeling on imbalanced domains," *ACM Computing Surveys (CSUR)*, vol. 49, no. 2, p. 31, 2016.

[8] Y.-A. Chung, H.-T. Lin, and S.-W. Yang, "Cost-aware pretraining for multiclass cost-sensitive deep learning," 2015.

[9] 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*, 2016, pp. 5375–5384.

[10] S. Wang, W. Liu, J. Wu, L. Cao, Q. Meng, and P. J. Kennedy, "Training deep neural networks on imbalanced data sets," in *2016 international joint conference on neural networks (IJCNN)*. IEEE, 2016, pp. 4368–4374.

[11] Y.-X. Wang, D. Ramanan, and M. Hebert, "Learning to model the tail," in *Advances in Neural Information Processing Systems*, 2017, pp. 7029–7039.

[12] S. Ando and C. Y. Huang, "Deep over-sampling framework for classifying imbalanced data," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2017, pp. 770–785.

[13] Y.-H. Liu, C.-L. Liu, and S.-M. Tseng, "Deep discriminative features learning and sampling for imbalanced data problem," in *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2018, pp. 1146–1151.

[14] S. S. Mullick, S. Datta, and S. Das, "Generative adversarial minority oversampling," *arXiv preprint arXiv:1903.09730*, 2019.

[15] Z. Li, Y. Jin, Y. Li, Z. Lin, and S. Wang, "Imbalanced adversarial learning for weather image generation and classification," in *2018 14th IEEE International Conference on Signal Processing (ICSP)*. IEEE, 2018, pp. 1093–1097.

[16] G. Douzas and F. Bacao, "Effective data generation for imbalanced learning using conditional generative adversarial networks," *Expert Systems with applications*, vol. 91, pp. 464–471, 2018.

[17] G. Mariani, F. Scheidegger, R. Istrate, C. Bekas, and C. Malossi, "Bagan: Data augmentation with balancing gan," *arXiv preprint arXiv:1803.09655*, 2018.

[18] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf, "Wasserstein auto-encoders," *arXiv preprint arXiv:1711.01558*, 2017.

[19] Q. Dong, S. Gong, and X. Zhu, "Imbalanced deep learning by minority class incremental rectification," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[20] 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*, 2014, pp. 2672–2680.

[21] E. Montahaei, M. Ghorbani, M. S. Baghshah, and H. R. Rabiee, "Adversarial classifier for imbalanced problems," *arXiv preprint arXiv:1811.08812*, 2018.

[22] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[23] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[24] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[25] G. K. Dziugaite, D. M. Roy, and Z. Ghahramani, "Training generative neural networks via maximum mean discrepancy optimization," 2015, pp. 258–267.

[26] Y. Jian, "Generalized categorization axioms," *arXiv preprint arXiv:1503.09082*, 2015.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014.

[28] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner *et al.*, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[29] H. Xiao, K. Rasul, and R. Vollgraf. (2017) Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms.

[30] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Citeseer, Tech. Rep., 2009.

[31] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," 2011.

[32] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[33] 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*, 2016, pp. 5375–5384.

[34] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classifiction tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[36] 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*. JMLR. org, 2017, pp. 2642–2651.

[37] S. Santurkar, L. Schmidt, and A. Madry, "A classification-based study of covariate shift in gan distributions," *arXiv preprint arXiv:1711.00970*, 2017.

[38] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.