Quality-Aware Self-Training on Differentiable Synthesis of Rare Relational Data

Chongsheng Zhang[1], **Yaxin Hou**[1], **Ke Chen**[2,3,*], **Shuang Cao**[1], **Gaojuan Fan**[1], **Ji Liu**[4]

[1]Henan University
[2]South China University of Technology
[3]Peng Cheng Laboratory
[4]Baidu Research
cszhang@ieee.org, {houyx, cscao, fangaojuan}@henu.edu.cn, chenk@scut.edu.cn, jiliuwork@gmail.com

Abstract

Data scarcity is a very common real-world problem that poses a major challenge to data-driven analytics. Although a lot of data-balancing approaches have been proposed to mitigate this problem, they may drop some useful information or fall into the overfitting problem. Generative Adversarial Network (GAN) based data synthesis methods can alleviate such a problem but lack of quality control over the generated samples. Moreover, the latent associations between the attribute set and the class labels in a relational data cannot be easily captured by a vanilla GAN. In light of this, we introduce an end-to-end self-training scheme (namely, Quality-Aware Self-Training) for rare relational data synthesis, which generates labeled synthetic data via pseudo labeling on GAN-based synthesis. We design a semantic pseudo labeling module to first control the quality of the generated features/samples, then calibrate their semantic labels via a classifier committee consisting of multiple pre-trained shallow classifiers. The high-confident generated samples with calibrated pseudo labels are then fed into a semantic classification network as augmented samples for self-training. We conduct extensive experiments on 20 benchmark datasets of different domains, including 14 industrial datasets. The results show that our method significantly outperforms state-of-the-art methods, including two recent GAN-based data synthesis schemes. Codes are available at https://github.com/yaxinhou/QAST.

Introduction

Data scarcity poses a great challenge to the analysis and learning of data in many different domains and applications. It is often associated with class-imbalanced distributions, in which the rare classes have significantly inadequate sample size, while the majority classes have an overly large amount of samples that may dominate the machine learning process. Compared to unstructured (image) data (Li, Kamnitsas, and Glocker 2020; Yang et al. 2022c; Liu, Chen, and Jia 2022), relational data often consists of multiple continuous and discrete attributes (feature dimensions) with diverse modes (Xu et al. 2019). Moreover, dependencies between a subset of attributes and the class labels as well as correlations between the attributes commonly exist in relational data, making the synthesis of relational data very challenging.

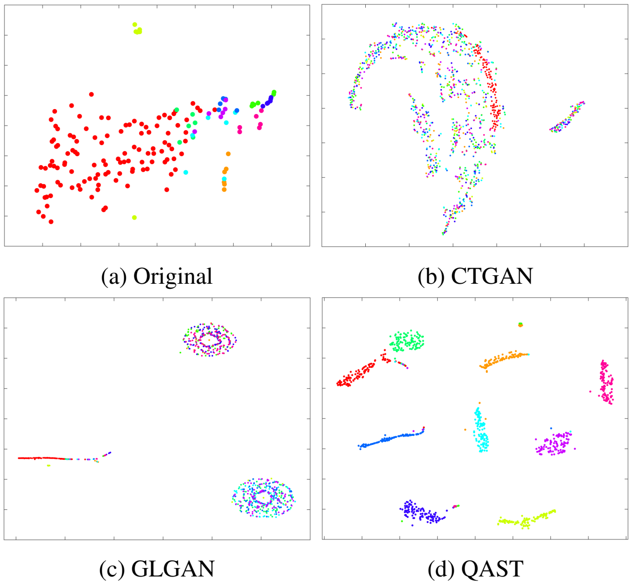
(a) Original
(b) CTGAN
(c) GLGAN
(d) QAST

Figure 1: Comparison of the projected feature distributions of the proposed QAST method with two comparative baselines CTGAN (Xu et al. 2019) and GLGAN (Wang et al. 2020), as well as the original data distribution, via T-SNE on the CWRU fault diagnosis dataset with an imbalance ratio of 1:20.

Existing approaches to the imbalance learning of relational data can be divided into two categories, which are cost-sensitive learning and data-balancing. The former attempts to improve the robustness of classifiers via introducing different penalties on the mis-classified samples in the cost functions, while the latter balances the number of samples across different classes to reduce the negative impact of imbalanced distributions, where data synthesis for the rare classes is an important and powerful option.

In the context of deep learning, data augmentation algorithms such as Mixup (Zhang et al. 2018) and its extensions can remarkably alleviate the suffer from data scarcity in the rare classes, but cannot ensure stable performance in view of the random operations in these augmentation methods. More importantly, these data augmentation methods rely on sim-

[*]Corresponding author.

ple linear operations such as interpolation and shuffle, therefore cannot avoid low-confident samples of rare classes off the manifolds. Inspired by recent success of differentiable rendering on the synthesis for semantic analysis such as unsupervised domain adaptation (Zhang et al. 2020), we aim to design an end-to-end learning based quality-aware self-training (QAST) scheme for the differentiable synthesis of rare relational data with high reliability and diversity.

Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) are a class of deep learning based generative methods that aim to learn the distribution of real data and create synthetic samples via the continuous competition between the generator and discriminator sub-networks. They are powerful in image generation, but their application to relational data is not straightforward (Xu et al. 2019). Although multiple tricks, *e.g.* mode-specific normalization and conditional generator, can be applied to relational data, the generated samples by GANs are typically lack of semantic labels, which therefore cannot be used to supervise deep models. Selection and annotation of high-confident generated samples should be a powerful option, which has been investigated in semi-supervised learning and unsupervised domain adaptation (Zhang et al. 2020; Zou et al. 2021; Chen et al. 2022). However, little attention has been devoted to the differentiable synthesis of rare relational data in the context of imbalance learning (Wang et al. 2020). Moreover, GANs cannot ensure that the generated data can preserve the strong and complicated correlations underlying in the attributes as well as the dependencies between a subset of attributes and the class labels that exist in the real relational data.

A Motivating Example. Consider a relational table with records of blood lipid levels of different patients. The attributes/features include <*Age, TC, TG, HDL-C, LDL-C, Apo A1, Apo B, Lp(a)*>, and the labels are {*Normal, Coronary Artery Disease, Stroke*}. Besides the normal values/ranges of the attributes, it presents the following associations: i) High levels of *TG* are typically accompanied with low levels of *HDL-C*; ii) *Apo A1* is positively correlated with *HDL-C*, and so is *Apo B* with *LDL-C*. Such inherent correlations between the features and the dependencies between a subset of features and the class labels are strong logical relationships that can not be easily identified and learned by a vanilla GAN.

To address the above issues, we propose QAST for the differentiable synthesis of rare relational data. Besides the ordinary generator-discriminator structure of a GAN, an extra classification branch is added in QAST for supervising the learning process, since high-quality synthetic samples are expected to promote the learning of a more powerful classification model. More importantly, we design a quality-aware semantic pseudo labeling module to control the quality of the generated samples and calibrate their semantic labels in a unified manner. In specific, the generator of a GAN can produce a large amount of unlabeled data (samples), we use the confidence of the discriminator on a synthetic sample to reflect its similarity to the real samples, and only the high-confident samples will be kept. For semantic pseudo labeling (Zou et al. 2021; Chen et al. 2022), our solution is simple yet effective: label quality is supported by a classifier committee consisting of multiple off-the-shelf pre-trained shallow classifiers which enable the representations of the relationships between the attribute set and the class labels from diverse perspectives, and the majority vote of class predictions is assigned as the semantic pseudo label to a high-confident sample.

In brief, in QAST, we first use the discriminator of a GAN to make a preliminary filtering on the quality of the generated features/samples. Next, in the semantic pseudo labeling module, we will check the dependencies between the features and the class labels in the generated samples using the classifier committee, which can further filter out less rational synthetic samples. For the remaining synthetic samples, we will calibrate their semantic labels via the majority vote by the classifier committee, then feed them into the semantic classifier as augmented samples for self-training. As shown in Figure 1, after generating samples for the rare classes using QAST, samples of different classes are more uniformly distributed in the feature space, and they can be more easily distinguished from those of other classes.

Main contributions of this work are summarized as follows:

- We propose a novel quality-aware self-training (QAST) scheme on the differentiable synthesis of rare relational data, which is a semantic task driven data generation approach to balance data distributions in an end-to-end learning manner.
- Technically, for selection and annotation of unlabeled synthetic samples, reliable samples with semantic pseudo labels calibrated via the majority vote of multiple pre-trained classifiers as geometric priors of feature distributions are selected for self-training.
- Results of extensive experiments on 20 benchmark datasets verify the effectiveness of our method, which consistently outperforms existing methods with significantly large margins.

Related Work

Imbalance Learning of Relational Data Existing approaches can be divided into cost-sensitive learning and data-balancing categories (Zhang et al. 2019a). Under-sampling and over-sampling are typical data balancing strategies for imbalance learning. Tomek link (Ivan 1976) and One-Sided Selection (OSS) (Kubat and Matwin 1997) are well-known under-sampling methods, while Random Over-Sampling (ROS) (Batista, Prati, and Monard 2004), SMOTE (Chawla et al. 2002), ADASYN (He et al. 2008), and MWMOTE (Barua et al. 2014) are representative over-sampling techniques. The Hellinger Distance Decision Trees (HDDT) (Cieslak et al. 2012) and EasyEnsemble (Liu, Wu, and Zhou 2009) are classical cost-sensitive learning algorithms for two-class imbalance learning, while (Zhou and Liu 2006) and (Murphey et al. 2007) are ensemble methods for multi-class imbalance learning. In recent years, many researchers have also explored using deep neural networks for imbalance classification, in which they often up-weight the rare classes or the difficult instances in the loss functions (Zhang et al. 2019b; Fernando and Tsokos 2022). Different

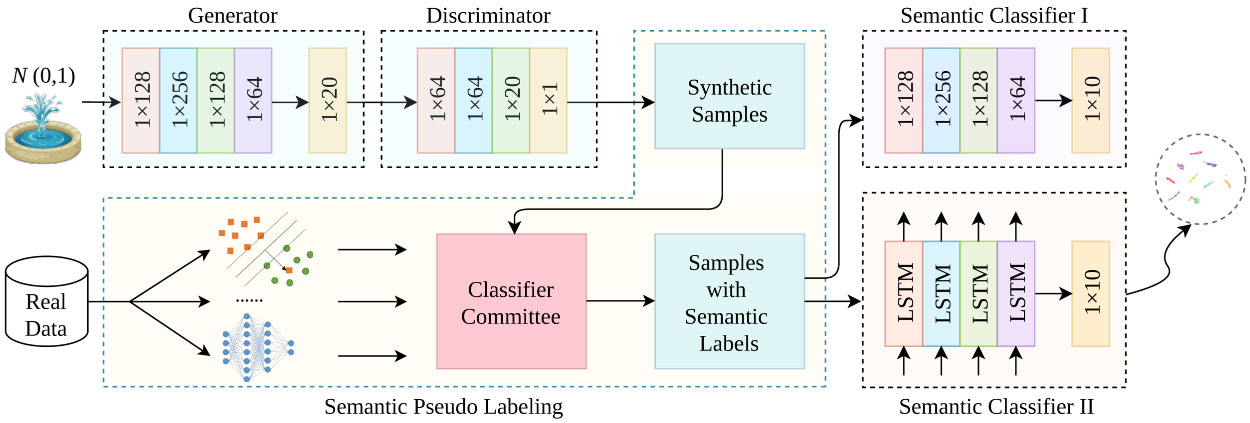

Figure 2: The architecture of our QAST scheme. It consists of three modules, *i.e.* a typical GAN for generating relational data, a quality-aware semantic pseudo annotator for synthetic samples, and a semantic classifier for multi-class imbalance classification. The first module is composed of the basic components of a GAN model, *i.e.* a generator and a discriminator. The second module first selects the high-confident samples then calibrates their semantic labels via the majority vote of multiple pre-trained shallow classifiers. The semantic classifier module first supervises the generator in sample generation, then builds the final classification model via a decoupled training process.

from the above approaches, we aim at designing an end-to-end self-training scheme for the differentiable synthesis of rare relational data with quality guarantees.

Deep Imbalance Learning of Image Data Deep imbalance learning is a research focus in recent years (Yang et al. 2022a). Existing approaches can also be divided into cost-sensitive learning and data-balancing categories. CB Loss (Cui et al. 2019) and LDAM (Cao et al. 2019) are typical cost-sensitive learning methods. The former is a class-wise re-weighting method that assigns weights to the prediction losses of samples using the inverse class frequency, while the latter modifies the Softmax loss function by adding a class-wise margin that reflects the class sparsity. The Gaussian Clouded Logit adjustment (GCL) loss (Li, Cheung, and Lu 2022) disturbs the margin with a random parameter to improve the robustness of the model. In (Li et al. 2022), a supervised contrastive learning method is designed that adds a loss term between the samples and the pre-defined class centers to regularize representation learning with geometric priors. CMO (Park et al. 2022) and RareGAN (Lin et al. 2022) are the latest data-balancing methods for deep imbalance learning. But their settings are different from ours, since we deal with labeled data and aim to improve the performance of imbalance learning through self-training.

Relational Data Synthesis Based on GAN The most related line of research to this work is GAN-based relational data synthesis. Table-GAN (Park et al. 2018) is the first attempt to generate relational data using deep learning, but its aim is to use the generated fake records to replace the original records for privacy concerns. The Conditional Tabular GAN (CTGAN) (Xu et al. 2019) exploits a conditional generator to generate relational data. The Medical Generative Adversarial Network (medGAN) (Choi et al. 2017) employs an auto-encoder structure to learn the salient features of discrete attributes in the embedding space. However, all the above methods have not investigated the synthesis of the rare classes in datasets with skewed distributions. GLGAN (Wang et al. 2020) employs both Auto-encoder and SMOTE to generate new samples with the consideration of local and global distributions, but the diversity of the generated samples is very limited due to the use of SMOTE. Overall, the feature quality and label semantics of the synthetic samples are uncontrolled in all the above methods, which may result in low-confident samples or noise labels that are detrimental to supervised learning.

Methodology

Given a training set of $\{(x_i, y_i)\}_{i=1}^{\mathrm{N}}$, where $x_i \in \mathcal{X}$ is an input relational sample and its corresponding class label is $y_i \in \mathcal{Y}$, N is the number of training instances, the problem of relational data classification aims to learn a mapping function $\Phi : \mathcal{X} \rightarrow \mathcal{Y}$ that classifies any test instance into one of the $\mathrm{K} = |\mathcal{Y}|$ object categories. Note that, the problem formulation for imbalanced data distributions is generally the same. The difference lies in the evaluation criteria, in which the performance on the rare classes is a critical concern. The mainstream deep learning based object classification algorithms can be formulated into a cascade of a feature extractor $\Phi_{\text{fea}} : \mathcal{X} \rightarrow \mathcal{F}$ and a classifier $\Phi_{\text{cls}} : \mathcal{F} \rightarrow \mathcal{Y}$ as

$$\Phi(x) = \Phi_{\text{cls}}(\boldsymbol{f}) \circ \Phi_{\text{fea}}(x) \tag{1}$$

where $\boldsymbol{f} \in \mathcal{F}$ denotes the feature output of $\Phi_{\text{fea}}(x)$.

Our end-to-end Quality-Aware Self-Training (QAST) scheme for rare relational data synthesis is composed of three modules, *i.e.* a shared GAN for generating relational data, a quality-aware semantic pseudo annotator for controlling both the feature quality and label quality of the generated samples, and a semantic classifier for multi-class imbalance classification, as depicted in Figure 2.

The shared GAN module consists of a generator following (Xu et al. 2019) to generate samples z to approach real samples' distribution to fool a discriminator for distinguishing real and synthetic samples, until reaching the Nash equilibrium. The semantic pseudo labeling module first removes a fraction of generated samples with low confidence in the binary classification of the discriminator, next assigns a majority vote of class predictions as the semantic pseudo label $\bar{y}$ for a selected high-confident synthetic sample $\bar{z}$, based on a classifier committee consisting of multiple pre-trained shallow classifiers.

Given augmented synthetic data $(\bar{z}, \bar{y})$ together with original real data (x, y), the semantic classifier module Φ is updated in a self-training style to guide data generation. A typical option is multiple fully-connected layers, *i.e.* the multiple-layer perceptron (MLP), which is adopted in the Phase 1 training of QAST in our experiments. During testing, each new sample is directly fed into the semantic classifier module Φ to predict the probabilities that it belongs to different classes.

GAN-Based Differentiable Synthesis

The differentiable synthesis of rare relational data in our QAST scheme is based on the popular GAN paradigm, which can produce a large amount of unlabeled data.

Generator of Relational Data The generator of a typical GAN desires diverse and reliable seeds as inputs to generate synthetic data of sufficient quality and diversity, which can be achieved via a random noise or SMOTE.

On one hand, randomness is introduced into the generation of seeds when using standard normal distributions, which can enrich the diversity of the seeds but also suffer from their unstable quality. On the other hand, SMOTE can generate more realistic samples from neighboring real samples of the same class, so the seeds generated via SMOTE are of high quality but are limited in diversity. Besides, Mixup (Zhang et al. 2018) creates a new synthetic image (sample) via the linear interpolation of two randomly selected images in both the feature space and label space, which can also be applied to the seed generation for GANs.

In our generator, we employ all the seed generation methods above to improve the quality and diversity of seeds generation for GANs. The rationale behind this design is as follows: if only random noise vector is used, the generator is very weak at the beginning, and the distribution of the synthetic samples encoded from the random noise vector using the generator is very different from the real data distribution, which are not conducive to the training of GANs. The addition of SMOTE and Mixup helps generate more reliable seeds, enabling the generator to quickly contact and learn the distribution of real data in the feature space, which should be beneficial in guiding the generator in encoding the seeds into more realistic data.

Such a generator is supervised by the the reconstruction loss $L_{rec}(X, Z)$ in (Rosca et al. 2017):

$$L_{rec}(X, Z) = \sum_{z \in Z, x \in X} ||z - x||^2 \tag{2}$$

where $X = \{x\}$ denotes the set of the normalized original samples by following (Xu et al. 2019), $Z = \{z\}$ denotes the set of synthetic samples created by the generator.

Discriminator The discriminator of a typical GAN is designed as a binary classification task to distinguish between synthetic and real samples, *i.e.* $\bar{X} = \{X \cup Z\}$. Following (Engelmann and Lessmann 2021), we employ a combination of the Binary Cross Entropy (BCE) loss and the Wassertein distance with gradient penalty as Eqn. (3):

$$L_D(\bar{X}) = L_{BCE}(\bar{X}) + L_{WGAN-GP}(\bar{X}) \tag{3}$$

where $L_D(\bar{X})$ is the discrimination loss, $L_{BCE}(\bar{X})$ denotes the BCE loss and $L_{WGAN-GP}(\bar{X})$ is the Wassertein distance with gradient penalty. $L_{BCE}(\bar{X})$ is depicted as

$$\begin{aligned} L_{BCE}(\bar{X}) = \sum_{x \in \bar{X}} &-[\mathbb{1}|_{x \in X} \log D(x) \\ &+ \mathbb{1}|_{x \in Z} \log(1 - D(x))] \end{aligned} \tag{4}$$

where $D(x)$ is the probability that a sample x belongs to the original data. $\mathbb{1}|_{x \in X}$ denotes the function is equal to 1 when the sample comes from the original data, while $\mathbb{1}|_{x \in Z}$ denotes the function is equal to 1 when the sample is synthesized by the generator. The Wassertein distance $L_{WGAN-GP}(\bar{X})$ is formulated as follows:

$$\begin{aligned} L_{WGAN-GP}(\bar{X}) = \mathbb{E}_{x \sim Z}[D(G(x))] - \mathbb{E}_{x \sim X}[D(x)] + \\ \lambda \mathbb{E}_{\hat{x} \sim P_{\hat{x}}}[(||\nabla_{\hat{x}} D(\hat{x})||_2 - 1)^2] \end{aligned} \tag{5}$$

where $\mathbb{E}_{x \sim Z}[D(G(x))] - \mathbb{E}_{x \sim X}[D(x)]$ is the original loss, and $\mathbb{E}_{\hat{x} \sim P_{\hat{x}}}[(||\nabla_{\hat{x}} D(\hat{x})||_2 - 1)^2]$ is a penalty loss by reducing the gradients of the discriminator to 1 using a combination of the normalized original sample x and the generated sample z, which is formulated as:

$$\tilde{x} = \varepsilon x + (1 - \varepsilon) z, \varepsilon \sim \text{Uniform}[0, 1], x \sim X, z \sim Z \tag{6}$$

where $\tilde{x}$ is a randomly averaged sample between x and z.

Quality-Aware Semantic Pseudo Labeling

Given a set of labeled real relational samples X and a set of unlabeled synthetic samples Z generated by the GAN-based differentiable synthesis, the problem of improving classification performance falls into the scope of typical semi-supervised learning. Self-training schemes (Zoph et al. 2020; Yang et al. 2022b) have verified their effectiveness in different semi-supervised learning problems, which assign semantic pseudo labels to unlabeled samples to complement the original data for more effective model training. QAST selects the high-confident synthetic samples that cannot be easily distinguished by the discriminator, using Eqn. (7):

$$Q_s(z) = \begin{cases} 1, D(z) \geqslant P_s \\ 0, \text{Otherwise} \end{cases} \tag{7}$$

where $D(z)$ denotes the probability that a synthetic sample is mis-classified as "real" by the discriminator, due to its realistic feature vectors, and P_s is the threshold. $Q_s(z) \geqslant P_s$ indicates that the sample $z \in Z$ is of high quality (i.e., with high confidence). When P_s is higher, fewer samples are selected, which sacrifices data synthesis efficiency for improving the feature quality of the samples, and vice-versa. In our approach, we set a self-adaptive P_s value that will gradually increase with the number of training epochs elapsed (denoted as i). Specifically, $P_s = (i + 50)/1000$.

For minimizing label noises of semantic pseudo annotation on synthetic relational samples, a number of off-the-shelf pre-trained shallow classifiers on real relational data are employed to form a classifier committee C_c, providing multi-view priors on relational data distribution and the logical dependencies between the attribute set and the class labels, which are therefore significantly more robust and reliable for semantic pseudo labeling than existing self-training schemes that rely on the predictions of a single integrated classifier. Specifically, when each classifier makes a probabilistic prediction on a sample, we count the majority vote of different classifiers (let n_p denote the vote count) to determine the final semantic pseudo label of the sample. To control the label quality, n_p should satisfy another self-adaptive threshold T_l, whereas samples with $n_p < T_l$ will be abandoned. The whole procedure is formulated as follows:

$$Q_l(z) = \begin{cases} 1, n_p \geq T_l \text{ and } T_l = \text{floor}(i \ / \ 100) \\ 0, \text{Otherwise} \end{cases} \tag{8}$$

where $Q_l(z) = 1$ indicates that sample z is of high label quality, floor$(\cdot)$ denotes the round down function, i is the index of training epochs. To balance between label quality control and curriculum learning, we set the value of T_l to evolve with the number of training epochs elapsed. Let $\bar{z}$ denote a selected synthetic sample, $Q_s(\bar{z}) \geqslant P_s$, let $\bar{y}$ be the class that receives the most votes (denoted as n_p) from the predictions of multiple pre-trained classifiers on $\bar{z}$, and $n_p \geq T_l$, then $\{(\bar{z}, \bar{y})\}$ will be the final synthetic sample to be fed into the semantic classifier module.

Self-Training for Semantic Classification

As a semantic task driven data synthesis method, QAST devises a semantic classifier module to train a classifier on both the real samples (x, y) and pseudo-labeled samples $(\bar{z}, \bar{y})$ to guide the generation of samples. The loss function for supervising semantic classification is presented in Eqn. (9) as:

$$L_C(\tilde{X}) = \sum_{(x,y)\in(\tilde{X})} L(x,y)|_{y\in MIC} + \varphi L(x,y)|_{y\in MAC} \tag{9}$$

where the data $\tilde{X}$ includes both the original samples and the generated ones. MIC denotes the set of the samples in the rare classes, while MAC denotes the set of the samples in the majority classes. φ is the weight to trade-off between the rare classes and majority classes. $L(\cdot)$ is the cross entropy loss as follows:

$$L(x,y) = -\sum_{c\in\mathcal{Y}} \log(\frac{e^{V(x,c)}}{\sum_{c'\in\mathcal{Y}} e^{V(x,c')}})\mathbb{1}|_{c=y} \tag{10}$$

where $V(x, c)$ is the logit element on class c for sample x, and $\mathbb{1}|_{c=y}$ denotes the function is equal to 1 when the sample belongs to class y.

To improve classification performance on the rare classes, we design a weighted Softmax loss to assign different weights to the loss of the *MIC* and *MAC* samples, respectively. The weight φ is inversely proportional to the number of samples in each class. Since the original dataset has a skewed data distribution, to balance the number of samples across classes, we set the total number of samples for each rare class (including both the real samples and the synthetic samples to be created by QAST) to be the averaged sample size of the majority classes.

End-To-End Model Training

For training the proposed model in an end-to-end style, the overall loss function is written as follows:

$$L_G \ = L_{rec}(Z, X) - L_D(\bar{X}) + L_C(\tilde{X}) \tag{11}$$

Since Z is expected to be very close to X, the reconstruction loss $L_{rec}(Z, X)$ should be small.

Similar to typical GANs, the generator of QAST is also expected to generate synthetic samples that can "fool" the discriminator, thus the loss function (11) has a minus discrimination loss.

The whole training procedure of QAST is divided into two decoupled training phases, *i.e.* the synthesis of high-confident samples with calibrated pseudo labels (Phase 1) and the final semantic classifier training (Phase 2). Within Phase 1, we focus on training a reliable generator and adopt a semantic classifier with moderate-depth (*i.e.* MLP). As with a vanilla GAN, the generator and discriminator are trained interactively. Moreover, we adopt the discriminator and a classifier committee of pre-trained shallow classifiers to filter out the out-of-distribution (low-confident) samples and less rational synthetic samples (if the majority vote n_p of a sample is smaller than T_l), respectively. The remaining synthetic samples will be assigned with labels via the majority vote of the classifier committee, which will then be combined with original real data for training a semantic classifier to supervise the generator in sample generation.

During Phase 2, we focus on deriving a powerful classification model. We first freeze the generator and discriminator networks, then replace the semantic classifier in Phase 1 (*i.e.* Semantic Classifier I) with a complex yet strong neural network (*i.e.* Semantic Classifier II) to build the final classification model. In our implementations, we adopt MLP for Semantic Classifier I and the BLSTM network (Sak, Senior, and Beaufays 2014) for Semantic Classifier II.

Note that typical GANs are often challenging to train, since they need careful regularization and expensive hyper-parameter sweeps (Sauer et al. 2021). In contrast, our QAST scheme follows a self-training style and the critical hyper-parameters are self-adaptive, which will be very easy to deploy and more applicable to real-world problems.

Experiments

Datasets and Settings

Datasets We evaluate comparative methods on 6 multi-class imbalanced datasets from the UCI repository. Moreover, we include 14 industrial datasets, which are the *CWRU Bearing* (Li et al. 2020) and *Gearbox fault* diagnosis datasets (Lin and Zuo 2003), and the NASA software defect repository

Methods	car				ecoli				glass			
	Acc	Macro-F1	G-Mean	pAR	Acc	Macro-F1	G-Mean	pAR	Acc	Macro-F1	G-Mean	pAR
ROS	96.06	91.98	93.89	70.54	81.36	74.17	71.91	20.97	67.45	56.10	24.06	11.71
SMOTE	95.11	89.05	92.05	69.95	81.30	75.48	73.42	23.18	67.08	59.61	52.13	15.40
BOS1	94.73	88.14	90.86	69.16	81.30	75.36	72.73	23.76	67.92	60.01	51.82	14.30
BOS2	88.12	75.09	84.58	67.40	79.82	73.73	70.78	23.32	66.51	54.13	18.41	6.40
SVM-SMOTE	95.50	86.18	87.31	66.28	82.10	75.21	72.50	22.92	65.85	58.36	51.28	17.66
ADASYN	96.54	90.69	91.54	69.00	82.53	76.27	73.41	23.10	65.47	58.13	51.14	17.66
CTGAN	79.92	60.21	74.62	58.44	82.59	75.64	71.87	22.26	67.08	57.84	53.03	16.26
GLGAN	94.43	84.68	86.47	62.20	82.82	77.35	75.27	24.78	70.23	62.12	53.69	14.77
QAST w/o ft	97.86	94.69	95.46	71.30	85.89	80.69	80.19	27.52	**74.77**	71.23	66.57	24.06
QAST	**98.15**	**94.96**	**96.30**	**72.79**	**86.50**	**84.70**	**85.25**	**31.14**	**74.77**	**71.55**	**69.78**	**30.05**
	pageblocks				satimage				thyroid			
	Acc	Macro-F1	G-Mean	pAR	Acc	Macro-F1	G-Mean	pAR	Acc	Macro-F1	G-Mean	pAR
ROS	94.88	68.09	33.12	34.57	89.30	87.50	86.91	39.93	98.44	91.86	93.71	69.25
SMOTE	95.25	73.86	66.47	34.60	89.62	87.87	87.30	40.08	98.44	91.86	93.71	69.25
BOS1	95.84	75.70	54.85	35.69	89.26	87.50	87.12	40.52	98.52	92.43	95.05	70.60
BOS2	95.07	71.70	60.87	34.46	88.95	87.12	86.97	40.62	96.38	81.70	87.55	64.45
SVM-SMOTE	94.77	67.28	34.78	34.52	89.38	87.65	87.06	39.33	98.13	91.05	92.31	66.26
ADASYN	95.14	70.42	58.83	35.66	89.21	87.51	87.09	39.36	98.16	91.27	92.70	66.54
CTGAN	94.55	67.95	61.29	35.69	89.03	86.87	85.79	38.64	93.40	75.09	85.96	64.18
GLGAN	95.07	74.47	70.13	35.42	89.88	87.72	86.20	39.08	98.41	90.59	91.23	67.78
QAST w/o ft	96.70	83.27	80.53	40.74	90.52	88.43	87.12	39.54	**99.03**	**93.49**	**97.53**	**73.44**
QAST	**97.07**	**85.96**	**81.25**	**44.99**	**91.17**	**89.60**	**89.33**	**42.20**	98.06	90.88	95.88	69.74

Table 1: Classification performance of comparative methods on 6 UCI datasets, in terms of Acc (%), Macro-F1 (%), G-Mean (%) and pAR (%). "QAST w/o ft" denotes the degenerated QAST without semantic classifier fine-tuning in Phase 2.

that contains 12 datasets (Shepperd et al. 2013). In the experiments, 10 cross-validations are adopted. Since there are no official splits for the above datasets, in each validation we randomly select half of the samples of each dataset as the training set, and use the other half as the test set.

Comparative Methods We compare QAST with eight baseline methods, including six well-established methods, which are ROS, SMOTE, BOS1 and BOS2 (Han, Wang, and Mao 2005), SVM-SMOTE (Nguyen, Cooper, and Kamei 2011), and ADASYN (He et al. 2008), and two recent GAN-based data synthesis schemes, *i.e.* CTGAN and GLGAN. For fair comparisons with the baselines on the quality of the generated samples, we degenerate our QAST without the semantic classifier fine-tuning in Phase 2 (*i.e.* QAST w/o ft), then let QAST w/o ft and each baseline method generate the same amount of samples, next adopt the same classification algorithm, *i.e.* gradient boosting decision tree (GBDT) (Friedman 2001; Zhang et al. 2019c), to build a specific classification model for each method and report their classification performance.

Performance Metrics The quantitative evaluation on the usability of the generated samples in supervised learning tasks is still an open question. Metrics such as the Inception Score (IS) (Salimans et al. 2016) and the Fréchet Inception Distance (FID) (Heusel et al. 2017) are just unsupervised metrics for evaluating the quality of the generated images. It would be ideal to use the rules/patterns underlying in the data (e.g., medical data), which are provided by experts, to evaluate the correlation among the attributes/dependencies in the generated relational data. Alternatively, classification accuracy as a surrogate can also reflect the quality/usability of the generated data. To this end, we adopt the commonly used metrics for imbalance classification (Zhang, Bi, and Soda 2017; Zhang et al. 2022), which are *Acc* (the overall accuracy), *Macro-F1* (macro-averaged F1 score) (Urbanowicz and Moore 2015), and *G-Mean* (the geometric mean of the per-class recall rate). Moreover, to reflect the specific performance on the rare classes, we also report the averaged per-class recall rate of the rare classes, denoted as *pAR*. In the empirical comparisons, we first use different generative methods to generate the same amount of synthetic samples. In specific, we set the total number of samples for each rare class (including both the real samples and the synthetic samples to be created by each method) to be the averaged sample size of the majority classes. Next, we use the same classification algorithm GBDT (or SVC/LinearSVC) to build a classification model upon the generated data (after combined with the original data) of each method, then compare the classification performance of each method on the same test set of each benchmark dataset which follows the same imbalanced distribution as the original data, to reflect the usability of the generated samples by each method.

Details about the Classifier Committee In our implementation, shallow classification algorithms in the scikit-learn library are adopted to train classification models with their default settings on the same training data to comprise the classifier committee. In total, we adopt *12* classification algorithms. Among them, there are *6* imbalance learning methods, which are *SMOTE* (Chawla et al. 2002), *ADASYN* (He et al. 2008), *SMOTEENN* and *SMOTETomek* (Batista, Bazzan, and Monard 2003), *EasyEnsemble* (Liu, Wu, and Zhou 2009), and *RUSBoost* (Seiffert et al. 2010). Meanwhile, we also include *6* ordinary classification algorithms, which are *SVC* and *LinearSVC* (Chang and Lin 2011), *MLP*, the Gra-

Datasets	Over-Sampling Methods			GAN-based Methods		
	SMOTE	BOS1	SVM-SMOTE	CTGAN	GLGAN	QAST
CWRU (IR=2)	93.10	93.16	93.80	94.45	93.04	**94.83**
CWRU (IR=5)	90.71	91.38	90.66	87.83	88.51	**94.53**
CWRU (IR=10)	86.65	89.02	86.36	81.62	87.97	**91.94**
CWRU (IR=20)	85.48	76.53	69.70	78.18	75.41	**91.58**
Gearbox (IR=2)	52.34	52.39	52.46	51.55	47.62	**53.59**
Gearbox (IR=5)	47.73	47.70	47.75	46.10	47.18	**49.50**
Gearbox (IR=10)	45.38	46.73	48.68	47.57	49.31	**50.80**
Gearbox (IR=20)	42.51	45.85	47.65	48.77	48.66	**49.57**
NASA CM1	55.35	54.06	57.31	54.48	57.54	**60.62**
NASA JM1	59.52	**59.88**	59.68	56.33	54.14	57.02
NASA KC1	62.65	60.50	63.39	60.19	60.32	**65.13**
NASA KC3	66.61	65.31	62.44	67.37	68.96	**72.61**
NASA MC1	57.55	57.81	58.97	61.74	64.39	**67.00**
NASA MC2	69.74	68.30	64.77	64.00	63.40	**72.35**
NASA MW1	66.84	66.67	65.91	61.37	62.94	**71.85**
NASA PC1	61.43	59.45	62.40	64.13	59.40	**71.44**
NASA PC2	54.82	56.92	56.92	58.79	57.63	**67.69**
NASA PC3	61.21	62.08	62.19	59.31	61.06	**65.86**
NASA PC4	75.85	75.40	73.98	73.37	71.27	**81.29**
NASA PC5	64.58	62.36	63.79	66.50	64.53	**68.58**

Table 2: The Macro-F1 (%) results of different methods on the CWRU Bearing and Gearbox Fault diagnosis datasets with varying imbalance ratios, and the NASA datasets. Due to space limit, here "QAST" denotes "QAST w/o ft".

dient Boosting Decision Tree (*GBDT*) (Friedman 2001), Logistic Regression (Hosmer Jr, Lemeshow, and Sturdivant 2013), and *BLSTM* (Sak, Senior, and Beaufays 2014). Each classifier committee is composed of *8* classification algorithms, half of which are randomly selected from the *6* imbalance learning methods and the other half come from the *6* ordinary classification algorithms. We will show that QAST is insensitive to the selection of the classification algorithms in the classifier committee.

Implementation Details In all the experiments, the Adam optimizer is used with a default learning rate of 0.0002, and the training epochs is set to 300 for all different methods. For fair comparisons, our QAST follows the baseline CTGAN to adopt PacGAN(Lin et al. 2018), which was designed to mitigate mode collapse.

Experimental Results

Evaluation on the UCI Datasets Table 1 presents the results of our QAST and other baseline methods. It can be observed that QAST outperforms all the baseline approaches on all the datasets, which confirms its effectiveness in imbalance classification. Specifically, the performance gain of QAST over the best competitor at each dataset are 5.67% on average in terms of Macro-F1, 7.35% on average in terms of G-Mean, and 5.7% on average in terms of pAR, all of which are very significant. These results support our motivation of self-training with high-confident synthetic samples with calibrated pseudo labels for improving the imbalance learning performance on the rare classes in datasets with skewed distributions. Besides, when comparing QAST with its degenerated QAST w/o ft, it is evident that the semantic classifier fine-tuning with a stronger network in Phase 2 is often beneficial to the performance enhancement of QAST.

Strategies	car		ecoli		glass	
	Macro-F1	pAR	Macro-F1	pAR	Macro-F1	pAR
$P_s \geq 0.1$	91.45	67.72	77.70	22.66	64.10	17.81
$P_s \geq 0.2$	89.90	64.40	69.29	20.39	59.79	11.37
$P_s \geq 0.3$	86.00	67.72	74.60	18.12	63.53	10.92
$T_l \geq 1$	90.90	66.88	78.13	27.43	60.86	11.56
$T_l \geq 2$	87.49	65.45	79.20	25.32	63.38	10.89
$T_l \geq 3$	88.85	68.57	77.87	24.30	61.62	11.56
Random	91.52	69.58	78.34	26.43	68.99	23.98
Mixup	91.29	66.88	76.45	22.95	59.18	10.50
SMOTE	90.59	64.60	74.46	22.74	62.71	11.76
Mixup+SMOTE	88.16	65.31	79.02	25.54	63.65	17.81
QAST w/o ft	**94.69**	**71.30**	**80.69**	**27.52**	**71.23**	**24.06**

Table 3: Ablation studies on the effects of different T_l and P_s settings on the Macro-F1 and pAR performance of QAST w/o ft, and comparisons on the performance of QAST when adopting different strategies for setting up generator seeds.

Evaluation on Multiple Industrial Datasets Results on industrial datasets, *i.e.* the *CWRU Bearing* and *Gearbox Fault* diagnosis datasets under different imbalance ratios (IR), and the NASA software defect repository, are shown in Table 2. Note that, for fair comparisons on the quality of the generated samples with the baselines, we only adopt QAST w/o ft. It is evident that QAST w/o ft can consistently outperform recent GAN-based synthesis, achieving the state-of-the-art performance on all the datasets (except the NASA JMI). The performance gain of QAST over the best competitor on each dataset is 3.39% on average in terms of Macro-F1, which demonstrates the effectiveness of our method when applying to real data of industrial applications.

Ablation Study

Impact of Self-adaptive Parameters in QAST In QAST, P_s and T_l are critical parameters for selecting high-confident samples and calibrating their semantic labels. In our design, both parameters will gradually increase with training epochs evolve, which will favour model training in a curriculum learning manner. In Table 3, experiments with different fixed values of P_s and T_l are conducted to compare with our QAST method on 3 UCI datasets. It is evident that, selecting high-confident samples with a fixed threshold P_s consistently yields inferior results than our self-adaptive mechanism. Similarly, the performance of semantic label calibration with a fixed threshold T_l often falls behind our self-adaptive mechanism as well. In the beginning, as the discriminator is relatively strong, it is highly likely that it can easily distinguish synthetic samples (recall that P_s denotes the probability threshold that the discriminator misclassifies a synthetic sample as "real"), which encourages a smaller P_s. When the generator is gradually improved during training, the probability that the discriminator misclassifies a synthetic sample increases, therefore a larger P_s is desired to keep more realistic samples. Similarly, with gradually increasing T_l during training, high-confident samples will be assigned with more reliable semantic labels. Overall, the self-adaptive parameter setting in QAST leads to high-quality synthetic samples with semantic labels.

Impact of the Hybrid Strategy for Setting Up Generator Seeds For setting up generator seeds, QAST combines three different approaches (Random, SMOTE and Mixup) to enrich the diversity of the input seeds. On one hand, the random mechanism generates seeds with good diversity, but the validity of the seeds cannot be guaranteed. On the other hand, SMOTE interpolates neighboring minority samples to generate more realistic samples, but suffers from the diversity problem, while Mixup interpolates the samples from different classes in both the feature space and label space, but cannot ensure a more balanced data distribution with the augmented samples. To integrate the strength of these strategies, QAST combines all of them to enhance the diversity of seeds for the generator. The third block in Table 3 reports the results of different generation strategies. It is clear that, the hybrid input seeds generation strategy for QAST achieves better performance than each single strategy.

Impact of the Selection of Classifier Committee on QAST To study the effects of the selection of classifier committee in the semantic pseudo labeling module, we compare the performance of QAST using the default classifier committee (*Classifier Group_1*) with another four groups of classifier combinations. For each competing classifier group, we randomly select 8 methods out of a set of 12 common classification algorithms, half of which are designed for imbalance classification. Details about these algorithms are provided in the supplementary file. We use GBDT to build a classification model over the real data and the synthetic data generated by the generator of QAST using each specific classifier committee. Table 4 reports the best performance among the baseline methods on each dataset (namely "Best Competitor"), then presents the results of QAST with five different groups of classifier committee. First of all, we observe that, irrespective of the selection of the classifier committee, our QAST consistently outperforms the best baseline methods on all the 3 datasets, and the improvement over the best competitor is 4.64% on average in terms of the Macro-F1, and 3.81% on average in terms of the pAR, both of which are significant. Second, comparing the performance of QAST with the other four groups of classifier committee, it is evident that their performance variance on each dataset is small. In summary, QAST is less sensitive to the selection of off-the-shelf classifiers for pseudo labeling.

Discussions

The novelty of QAST lies in utilizing an ensemble of model-driven shallow classifiers to supervise the data-driven self-supervised learning process for generating annotated synthetic samples. On one hand, decision boundaries of some shallow classifiers can intrinsically reveal feature distribution. On the other hand, all the classifiers aim to discover the inherent relations between the attributes and the class labels. As a result, using multiple pre-trained shallow classifiers for synthetic data annotation can benefit from an ensemble of geometric priors from different perspectives.

Strategies	car		ecoli		glass	
	Macro-F1	pAR	Macro-F1	pAR	Macro-F1	pAR
Best Competitor	91.98	70.54	77.35	24.78	62.12	17.66
Classifier Group_1	94.69	71.30	80.69	27.52	71.23	24.06
Classifier Group_2	93.39	70.54	79.43	28.22	72.06	29.99
Classifier Group_3	94.10	71.32	80.21	28.17	68.63	24.70
Classifier Group_4	94.26	70.54	78.66	27.80	72.08	23.39
Classifier Group_5	95.06	72.06	80.05	28.23	72.30	24.26

Table 4: Ablation studies on the influence of the choices of classifier committee on the performance of QAST w/o ft.

The main concern in this work lies in the generation and annotation of high-quality synthetic data for rare classes in an end-to-end learning manner, which are then combined with real data to construct a more balanced data set for semantic classification. To this end, our paper proposes an effective self-training scheme on the differentiable synthesis of rare classes with built-in quality control, while both modules of the GAN based data synthesis and pseudo labelling are not claimed as our contribution. Technically, considering lack of semantic labels of generated synthetic samples, reliable samples with pseudo labels calibrated via the majority voting of multiple pre-trained classifiers are selected for self-training, which are verified to achieve remarkable performance gain in our large-scale experiments.

Conclusions

In this work, we propose a general end-to-end learning scheme (QAST) on the differentiable synthesis of rare relational data to mitigate the data scarcity problem. QAST separates data synthesis into two perspectives: the discriminator of a GAN is used to control the quality of feature generation, while a semantic pseudo labeling module with a diverse committee of pre-trained classifiers is proposed to control the quality of label generation. We carry out extensive experiments on 20 benchmark datasets of different domains, which show that QAST consistently outperforms the baselines by a large margin. We also show that QAST is less sensitive to the selection of classifiers in the committee and demonstrates consistent advantages over the competitors. In future work, we will extend QAST to generate labeled images for data augmentation in vision tasks.

Acknowledgements

The first author is affiliated with and partially supported by Henan University, the Henan Provincial Key Laboratory of Big Data Analysis and Processing, the School of Computer and Information Engineering, the Laboratory of Yellow River Cultural Heritage, and the MOE Key Research Institute of Yellow River Civilization and Sustainable Development. We sincerely thank all the anonymous reviewers for their constructive suggestions and kind help.

References

Barua, S.; Islam, M. M.; Yao, X.; and Murase, K. 2014. MWMOTE–Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning. *IEEE*

Transactions on Knowledge and Data Engineering, 26(2): 405–425.

Batista, G.; Bazzan, A.; and Monard, M.-C. 2003. Balancing Training Data for Automated Annotation of Keywords: a Case Study. In *the Proc. of Workshop on Bioinformatics*, 10–18.

Batista, G.; Prati, R.; and Monard, M. 2004. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explorations*, 6(1): 20–29.

Cao, K.; Wei, C.; Gaidon, A.; Aréchiga, N.; and Ma, T. 2019. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. In *Advances in Neural Information Processing Systems (NeurIPS 2019)*, 1565–1576.

Chang, C.-C.; and Lin, C.-J. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3): 1–27.

Chawla, N.; Bowyer, K.; Hall, L.; and Kegelmeyer, W. 2002. SMOTE: Synthetic Minority Over-sampling TEchnique. *Journal of Artificial Intelligence Research*, 16: 321–357.

Chen, Y.; Wang, Z.; Zou, L.; Chen, K.; and Jia, K. 2022. Quasi-Balanced Self-Training on Noise-Aware Synthesis of Object Point Clouds for Closing Domain Gap. In *Proceedings of the European Conference on Computer Vision*, 728–745.

Choi, E.; Biswal, S.; Malin, B.; Duke, J.; Stewart, W. F.; and Sun, J. 2017. Generating Multi-label Discrete Patient Records using Generative Adversarial Networks. In *Proceedings of the 2nd Machine Learning for Healthcare Conference*, 286–305.

Cieslak, D. A.; Hoens, T. R.; Chawla, N. V.; and Kegelmeyer, W. P. 2012. Hellinger distance decision trees are robust and skew-insensitive. *Data Mining and Knowledge Discovery*, 24(1): 136–158.

Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; and Belongie, S. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition(CVPR 2019)*, 9268–9277.

Engelmann, J.; and Lessmann, S. 2021. Conditional Wasserstein GAN-based oversampling of tabular data for imbalanced learning. *Expert Systems with Applications*, 174: 114582.

Fernando, K. R. M.; and Tsokos, C. P. 2022. Dynamically Weighted Balanced Loss: Class Imbalanced Learning and Confidence Calibration of Deep Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(7): 2940–2951.

Friedman, J. H. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A. C.; and Bengio, Y. 2014. Generative Adversarial Nets. In *Neural Information Processing Systems (NIPS)*, 2672–2680.

Han, H.; Wang, W.; and Mao, B. 2005. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In *2005 International Conference on Intelligent Computing (ICIC05)*, volume 3644 of *Lecture Notes on Computer Science*, 878–887.

He, H.; Bai, Y.; Garcia, E. A.; and Li, S. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *IEEE International Joint Conference on Neural Networks*, 1322–1328.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Annual Conference on Neural Information Processing Systems (NeurIPS 2017)*, 6626–6637.

Hosmer Jr, D. W.; Lemeshow, S.; and Sturdivant, R. X. 2013. *Applied logistic regression*, volume 398. John Wiley & Sons.

Ivan, T. 1976. Two Modifications of CNN. *IEEE Transactions on Systems, Man, and Cybernetics*, 7(2)(11): 769–772.

Kubat, M.; and Matwin, S. 1997. Addressing the curse of imbalanced training sets: one-sided selection. In *14th International Conference on Machine Learning (ICML97)*, 179–186.

Li, M.; Cheung, Y.-m.; and Lu, Y. 2022. Long-tailed Visual Recognition via Gaussian Clouded Logit Adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6929–6938.

Li, T.; Cao, P.; Yuan, Y.; Fan, L.; Yang, Y.; Feris, R. S.; Indyk, P.; and Katabi, D. 2022. Targeted supervised contrastive learning for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6918–6928.

Li, Y.; Wang, X.; Si, S.; and Huang, S. 2020. Entropy Based Fault Classification Using the Case Western Reserve University Data: A Benchmark Study. *IEEE Transactions on Reliability*, 69(2): 754–767.

Li, Z.; Kamnitsas, K.; and Glocker, B. 2020. Analyzing overfitting under class imbalance in neural networks for image segmentation. *IEEE Transactions on Medical Imaging*, 40(3): 1065–1077.

Lin, J.; and Zuo, M. 2003. Gearbox fault diagnosis using adaptive wavelet filter. *Mechanical systems and signal processing*, 17(6): 1259–1269.

Lin, Z.; Khetan, A.; Fanti, G.; and Oh, S. 2018. PacGAN: The power of two samples in generative adversarial networks. In *Advances in Neural Information Processing Systems*, 1505–1514.

Lin, Z.; Liang, H.; Fanti, G.; and Sekar, V. 2022. RareGAN: Generating Samples for Rare Classes. In *Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI 2022)*, 7506–7515.

Liu, K.; Chen, K.; and Jia, K. 2022. Convolutional Fine-Grained Classification With Self-Supervised Target Relation Regularization. *IEEE Transactions on Image Processing*, 31: 5570–5584.

Liu, X.-Y.; Wu, J.; and Zhou, Z.-H. 2009. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2): 539–550.

Murphey, Y. L.; Wang, H.; Ou, G.; and Feldkamp, L. A. 2007. OAHO: an effective algorithm for multi-class learning from imbalanced data. In *2007 International Joint Conference on Neural Networks*, 406–411.

Nguyen, H. M.; Cooper, E. W.; and Kamei, K. 2011. Borderline over-sampling for imbalanced data classification. *Int. Journal of Knowledge Engineering and Soft Data Paradigms*, 3(1): 4–21.

Park, N.; Mohammadi, M.; Gorde, K.; Jajodia, S.; Park, H.; and Kim, Y. 2018. Data Synthesis Based on Generative Adversarial Networks. *Proceeding of the VLDB Endowment*, 11(10): 1071–1083.

Park, S.; Hong, Y.; Heo, B.; Yun, S.; and Choi, J. Y. 2022. The Majority Can Help The Minority: Context-rich Minority Oversampling for Long-tailed Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6887–6896.

Rosca, M.; Lakshminarayanan, B.; Warde-Farley, D.; and Mohamed, S. 2017. Variational approaches for auto-encoding generative adversarial networks. *arXiv preprint arXiv:1706.04987*.

Sak, H.; Senior, A.; and Beaufays, F. 2014. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *arXiv preprint arXiv:1402.1128*.

Salimans, T.; Goodfellow, I. J.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved Techniques for Training GANs. In *Annual Conference on Neural Information Processing Systems (NeurIPS 2016)*, 2226–2234.

Sauer, A.; Chitta, K.; Müller, J.; and Geiger, A. 2021. Projected GANs Converge Faster. In *Annual Conference on Neural Information Processing Systems (NeurIPS 2021)*, 17480–17492.

Seiffert, C.; Khoshgoftaar, T. M.; Hulse, J. V.; and Napolitano, A. 2010. RUSBoost: A Hybrid Approach to Alleviating Class Imbalance. *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, 40(1): 185–197.

Shepperd, M.; Song, Q.; Sun, Z.; and Mair, C. 2013. Data Quality: Some Comments on the NASA Software Defect Datasets. *IEEE Transactions on Software Engineering*, 39(9): 1208–1215.

Urbanowicz, R. J.; and Moore, J. H. 2015. ExSTraCS 2.0: description and evaluation of a scalable learning classifier system. *Evolutionary intelligence*, 8(2): 89–116.

Wang, W.; Wang, S.; Fan, W.; Liu, Z.; and Tang, J. 2020. Global-and-Local Aware Data Generation for the Class Imbalance Problem. In *Proceedings of the 2020 SIAM International Conference on Data Mining (SDM 2020)*, 307–315.

Xu, L.; Skoularidou, M.; Cuesta-Infante, A.; and Veeramachaneni, K. 2019. Modeling Tabular data using Conditional GAN. In *Advances in Neural Information Processing Systems (NIPS)*, 7333–7343.

Yang, L.; Jiang, H.; Song, Q.; and Guo, J. 2022a. A Survey on Long-Tailed Visual Recognition. *International Journal of Computer Vision*, 130(7): 1837–1872.

Yang, L.; Zhuo, W.; Qi, L.; Shi, Y.; and Gao, Y. 2022b. St++: Make self-training work better for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022)*, 4268–4277.

Yang, X.; Wang, Y.; Chen, K.; Xu, Y.; and Tian, Y. 2022c. Fine-Grained Object Classification via Self-Supervised Pose Alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7399–7408.

Zhang, C.; Bi, J.; and Soda, P. 2017. Feature selection and resampling in class imbalance learning: Which comes first? An empirical study in the biological domain. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2017)*, 933–938.

Zhang, C.; Bi, J.; Xu, S.; Ramentol, E.; Fan, G.; Qiao, B.; and Fujita, H. 2019a. Multi-Imbalance: An open-source software for multi-class imbalance learning. *Knowledge Based Systems*, 174: 137–143.

Zhang, C.; Soda, P.; Bi, J.; Fan, G.; Almpanidis, G.; Garcia, S.; and Ding, W. 2022. An empirical study on the joint impact of feature selection and data resampling on imbalance classification. *Applied Intelligence*, 1–13.

Zhang, C.; Tan, K. C.; Li, H.; and Hong, G. S. 2019b. A Cost-Sensitive Deep Belief Network for Imbalanced Classification. *IEEE Transactions on Neural Networks and Learning Systems*, 30(1): 109–122.

Zhang, C.; Zhang, Y.; Shi, X.; Almpanidis, G.; Fan, G.; and Shen, X. 2019c. On Incremental Learning for Gradient Boosting Decision Trees. *Neural Processing Letters*, 50(1): 957–987.

Zhang, H.; Cissé, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond Empirical Risk Minimization. In *6th International Conference on Learning Representations (ICLR 2018)*.

Zhang, Y.; Deng, B.; Jia, K.; and Zhang, L. 2020. Label Propagation with Augmented Anchors: A Simple Semi-supervised Learning Baseline for Unsupervised Domain Adaptation. In *16th European Conference on Computer Vision (ECCV 2020)*, 781–797.

Zhou, Z.; and Liu, X. 2006. On Multi-Class Cost-Sensitive Learning. In *The Twenty-First Conference on Artificial Intelligence (AAAI)*, 567–572.

Zoph, B.; Ghiasi, G.; Lin, T.; Cui, Y.; Liu, H.; Cubuk, E. D.; and Le, Q. 2020. Rethinking Pre-training and Self-training. In *Annual Conference on Neural Information Processing Systems (NeurIPS 2020)*, 3833–3845.

Zou, L.; Tang, H.; Chen, K.; and Jia, K. 2021. Geometry-aware self-training for unsupervised domain adaptation on object point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6403–6412.