Interpretable Minority Synthesis for Imbalanced Classification

Yi He , **Fudong Lin** , **Xu Yuan*** and **Nian-Feng Tzeng**
University of Louisiana at Lafayette
{yi.he1, fudong.lin1, xu.yuan, nianfeng.tzeng}@louisiana.edu

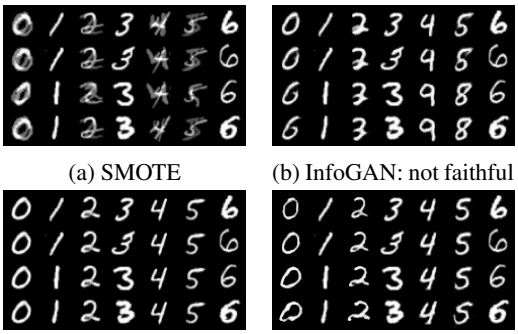
Abstract

This paper proposes a novel oversampling approach that strives to balance the class priors with a considerably imbalanced data distribution of high dimensionality. The crux of our approach lies in learning *interpretable* latent representations that can model the synthetic mechanism of the minority samples by using a generative adversarial network (GAN). A Bayesian regularizer is imposed to guide the GAN to extract a set of salient features that are either disentangled or intensionally entangled, with their interplay controlled by a prescribed structure, defined with human-in-the-loop. As such, our GAN enjoys an improved sample complexity, being able to synthesize high-quality minority samples even if the sizes of minority classes are extremely small during training. Empirical studies substantiate that our approach can empower simple classifiers to achieve superior imbalanced classification performance over the state-of-the-art competitors and is robust across various imbalance settings. Code is released in github.com/fudonglin/IMSIC.

1 Introduction

Imbalanced data abound in a variety of human endeavors such as online transactions, biomedical images, social media, among many others. In such data, whereas the distribution of samples representing the majority and minority classes is highly skewed, the minority samples are usually of much greater interest and tend to incur huge costs if misclassified. For example, misclassifying a credit card fraud from millions of legitimate transactions would result in a financial loss, while failing to diagnose lung cancer from a massive amount of CT images may end up with fatalities.

However, building classifiers from imbalanced data remains a challenging task, where the key lies in how to balance the class priors so that the minority samples are more focused. To that end, oversampling is a generic and preferred solution over other methods such as undersampling or cost-sensitive learning, as it neither suffers from the removal of valuable information nor entails expert knowledge to craft domain-specific objective functions [Branco *et al.*, 2016].



(a) SMOTE (b) InfoGAN: not faithful

(c) InfoGAN: not diverse (d) Our Approach

Figure 1: Illustration of the outputs of linear and deep oversampling techniques, where the digits "0", "2", "4", and "5" represent the minority classes in an imbalanced ratio of $1:100$. (a) SMOTE [Chawla *et al.*, 2002] outputs noises erroneously as non-existent digits. InfoGAN [Chen *et al.*, 2016] may either (b) lose track of digit identities so as to generate samples from wrong classes or (c) overfit to identical visual patterns thereby lacking diversity. (d) Our approach synthesizes samples being both faithful and diverse.

Yet, the crux of oversampling relies on two principles: i) the *fidelity*, where synthesized samples should truly belong to the minority classes, and ii) the *diversity*, where synthesized samples should carry a variety of patterns so as to better the classifier. To implement the two principles, the dominant paradigm is to uncover the geometric shape underlying data, on which new data points are synthesized within the local regions, comprising both majorities and minorities (so to cover various patterns), with the minority class dominating (so the new points are likely to be the true minorities).

Unfortunately, this conventional oversampling paradigm only works well for low-dimensional data and fails to generalize to high-dimensional and complex data, such as images. The reason is that the high-dimensional vectors usually lie on or close to a non-linear manifold that is described by a much smaller number of latent features. The previous methods applying a linear combination of existing data to synthesize new points fail to respect such non-linear structures. As a result, the synthesized samples being disparate from the manifold will become *noises*, not belonging to any class, as shown in

*Corresponding author: Dr. Xu Yuan (xu.yuan@louisiana.edu)

Figure 1a, defying the fidelity principle.

To overcome the issue, it seems plausible to leverage deep generative models to capture the non-linear intrinsic structures of data, hoping that the new points are synthesized faithfully. This idea, however, cannot work well, as the deep models are usually data-hungry, while the minority classes can afford very limited numbers of samples only. In effect, the use of deep generative models gives rise to a prominent fidelity-diversity tradeoff. That is, on the one hand, if the minority samples do cover a variety of patterns (*e.g.*, visual concepts such as rotation and scaling), the samples carrying each pattern cannot suffice to train a robust deep model, where the synthesized samples are likely to be noises (*e.g.*, images with non-existent or wrong digit identities) as shown in Figures 1b. On the other hand, if the minority samples carry limited patterns only, the deep models tend to discover a converged region, making the synthesized samples nearly-identical as shown in Figures 1c. The trained classifier would suffer from overfitting and hence generalize poorly to the test data that carry additional or different patterns.

To tackle the undesired tradeoff between fidelity and diversity, this paper proposes a novel generative adversarial network (GAN) architecture that learns *interpretable* latent representations to describe the non-linear structures underlying data. Our key idea is to impose a Bayesian regularizer to encourage the extracted features to interact in accordance with a dependency structure given a priori. This structure can guide the GAN to discover a set of features that carry salient semantic meanings as we desire by controlling over the granularity of the latent space. In particular, on the one hand, we enforce the salient features that determine the class membership of synthesized samples be strictly disentangled from other features, so as to practice the fidelity principle. On the other hand, we allow the features representing visual concepts to be entangled in an informative manner, so as to observe the diversity principle. As such, our GAN architecture enjoys better sample efficiency and thus can synthesize high-quality minority samples across various imbalance settings.

Specific contributions of this paper are as follows.

1. We propose a novel oversampling approach by leveraging GAN to model the data generating mechanism with non-linear latent representations.
2. A Bayesian regularizer is crafted to impose structural constraints upon the learned space, guiding GAN to discover a set of salient features that convey semantic meanings and form dependency structures in a human-controlled, and hence interpretable, manner.
3. Extensive experiments are carried out, and the results substantiate that our approach can enable simple classifiers to outperform the state-of-the-art imbalanced classification competitors across various imbalanced ratios.

2 Preliminaries, Challenge, and Our Idea

Problem Statement. Consider a set of N samples $\{\mathbf{x}_i\}_{i=1}^N \in \mathcal{X}$, deemed as the *real* data, which are drawn from a complex yet unknown distribution $\mathbf{x}_i \sim \mathbb{P}_{real}$. Suppose they are in C classes, with the numbers of samples in each class being $\{n_1, \ldots, n_C\}$ and $N = \sum_{c=1}^{C} n_c$. In this paper, we consider a highly imbalanced setting, where the size of the largest class (*i.e.*, the majority) is much larger than that of the smallest one (*i.e.*, the minority) by at least two orders of magnitudes. Our goal is to synthesize new samples for minority classes, balancing the class priors for accurate classification.

We frame our design in a generative adversarial network (GAN) regime. Let G and D denote the generator and the discriminator, respectively. For a conventional GAN, G takes in a noise vector $\mathbf{z} \in \mathbb{R}^{d_z}$ and outputs the synthesized, *fake* data; then, D receives both the data generated by G and the original samples, striving to classify them as real or fake. As such, the training process of a GAN can be deemed as a two-player *minimax* game between G and D, defined as follows.

$$\min_G \max_D \mathcal{L}_{\text{adv}}(G, D) - \lambda\Omega(G), \tag{1}$$

where $\mathcal{L}_{\text{adv}}(\cdot, \cdot)$ denotes the adversarial loss [Goodfellow *et al.*, 2014] and $\Omega(G)$ is the regularization term. The intuition behind Eq. (1) is that, while G aims to generate seemingly authentic data to fool D, D tries its best to screen the fake data out. To this end, subsequent GAN studies devise different regularizers to impose constraints on the searching space of G, hoping the learned G to possess various properties.

InfoGAN for Salient Feature Disentanglement. To have control over the patterns of synthesized data, InfoGAN [Chen *et al.*, 2016] tailored a information-theoretic regularizer that maximizes mutual information between a set of observed variables and the generator distribution. Specifically, G receives the input vectors in the form of $[\mathbf{z}, \mathbf{c}]^\top$, where $\mathbf{c} \in \mathbb{R}^{d_c}$ is a set of observed variables.

The regularizer is defined as $\Omega(G) = I\big(\mathbf{c}; G(\mathbf{z}, \mathbf{c})\big)$, where $I(\cdot, \cdot)$ calculates mutual information and $G(\mathbf{z}, \mathbf{c})$ is the fake samples output by G. The intuition behind this design is to preserve the information of observed variables during the generation process, so they can convey certain semantic meanings, where each variable becomes an *disentangled* salient feature of the synthesized samples (*e.g.*, digit identities and visual concepts of images).

Challenge: the Fidelity-Diversity Tradeoff. To adopt InfoGAN (or other GAN variants) for the minority synthesis, a prominent tradeoff between fidelity and diversity arises. The reason is that, as the mutual information term is computationally intractable, InfoGAN casts it into a variational lower bound maximization problem instead. This approximation, however, makes the optimization problem data-hungry, while in our setting the minority classes can afford very limited samples only. Hence, the independence of the extracted latent features cannot be guaranteed, such that the generated samples are likely to be noises and hence are not *faithful*. For example in MNIST, the synthesized minorities may be non-existent digits or seem to be the digits from majority classes.

A naïve solution to counter this fidelity issue is to replicate or replay the minority samples multiple times along with the majority samples during training. Unfortunately, this solution leads to *overfitting* since the minority classes usually contain samples having limited visual patterns. As a result, the synthesized minorities are likely to be identical, lacking

diversity. A classifier trained on such synthesized minorities cannot generalize well to the test data with additional or different patterns, yielding inferior classification performance.

Our Idea: Bayesian Network Regularizer. To improve sample efficiency, we model the joint distribution of the observed variables and the extracted latent features with a Bayesian network [Heckerman and Wellman, 1995; Beyazit *et al.*, 2020], which in turn functions as a GAN regularizer. Three observations motivate this idea. *First*, a Bayesian structure captures feature relationships, leading to the extraction of latent representation in a finer level of granularity, which allows to model the joint distribution of variables in a high degree of freedom. Namely, it can control the structure of the latent feature space, in a way that it lets some variables be strictly disentangled from other variables, while allowing some other variables to be entangled in a pre-defined manner. *Second*, compared to the unstructured method, having control over the feature dependency structure lifts the requirement of massive amounts of training data, making it better suit for class-imbalanced settings. *Third*, capturing the dependency structure among the latent features provides additional understanding about the data. Such an understanding in turn fosters the (deep) model interpretability.

3 Our Approach

We aim to learn interpretable minority representations via the joint consideration of strictly disentangled and informatively entangled latent features. In Section 3.1, we elaborate the learning objective function and the intuition behind its design. In Section 3.2, we scrutinize the parameter estimation strategy tailored for efficiency improvement.

3.1 Bayesian Regularizer for Minority Synthesis

Our GAN architecture is illustrated in Figure 2. Let $[\mathbf{z}, \mathbf{c}]^\top$ denote the input variables, in which the noise vector $\mathbf{z}$ implicitly approximates the sample probability distribution and $\mathbf{c}$ represents the observed variables expected to take on salient meanings after training. Without loss of generality, we follow [Chen *et al.*, 2016; Beyazit *et al.*, 2020] to draw $\mathbf{z}$ and $\mathbf{c}$ from factored distributions such as Gaussian with identity covariance, *i.e.*, $\mathbf{z}, \mathbf{c} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. θ_G and θ_D parameterize the generator G and discriminator D, respectively.

Latent Code Extraction. We consider D to comprise two sub-networks, namely, a feature extractor and a classifier. In particular, after receiving a sample being fake (*i.e.*, generated by G) or real, a feature extractor F parameterized by θ_F extracts the latent features from the sample, defined as:

$$F\left(\left\{\mathbf{x}, G([\mathbf{z}, \mathbf{c}]^\top, \theta_G)\right\}; \theta_F\right) = [\mathbf{z}', \mathbf{c}']^\top, \tag{2}$$

where the dimensions of $\mathbf{z}'$ and $\mathbf{c}'$ are identical to those of $\mathbf{z}$ and $\mathbf{c}$, respectively. Here, we deem $\mathbf{c}'$ as the *latent code*, in which each variable represents a salient feature, forming a dependency structure with the observed variables in $\mathbf{c}$.

Bayesian Modeling. We let a Bayesian network $\mathcal{B}$ model the joint distribution of $\mathbf{c}$ and $\mathbf{c}'$, *w.r.t.* a given connectivity pattern, as a product of local distribution probabilities. Denoted by $\mathbf{p}_i = \{p_{i1}, p_{i2}, \ldots, p_{ik}\}$ the *parents* of the i-th variable c'_i in the latent code $\mathbf{c}'$, $k < |\mathbf{c}| + |\mathbf{c}'|$. Note, 1) the observed variables in $\mathbf{c}$ do not have parents as they are sampled directly from a Gaussian; 2) any latent c'_i can be parented by both observed or latent variables according to the topology of $\mathcal{B}$. The local conditional probability of c'_i is defined as:

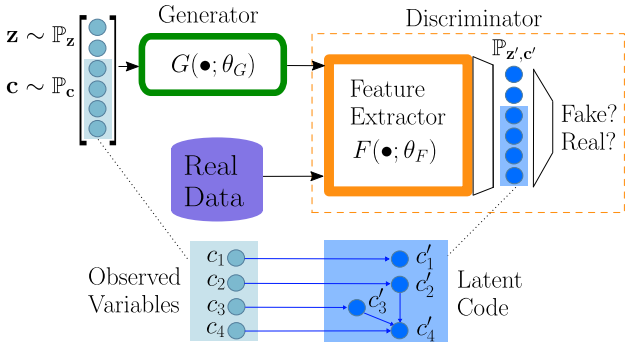


Figure 2: The architecture of our proposed GAN with the dependency structure among the observed variables and the extracted latent code represented by a Bayesian network to guide the GAN training. Here, four salient features are extracted, among which 1) c'_1 is strictly disentangled and 2) c'_2 and c'_3 are entangled with c'_4.

$$\begin{aligned}\mathbb{P}(c'_i \mid \mathbf{p}_i) &= \mathcal{N}(\mathbf{w}_i^\top \mathbf{p}_i; \sigma_i^2) \\ &= \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{(c'_i - \mathbf{w}_i^\top \mathbf{p}_i)^2}{2\sigma_i^2}\right],\end{aligned} \tag{3}$$

where $\mathbf{w}_i$ and σ_i^2 denote the weight vector representing a linear structure between c'_i and its parents and the variance capturing the Gaussian noise around such structure, respectively.

The optimal parameters $(\mathbf{w}_i, \sigma_i^2)$ for Eq. (3) can be estimated in each observed training batch. Specifically, in a batch of size m, we denote the values of the i-th latent variable c'_i as $\mathbf{c}'_i := [c'_{i,1}, \ldots, c'_{i,m}]^\top \in \mathbb{R}^m$ and the values of its parents as $\mathbf{P}_i := [\mathbf{p}_{i,1}, \ldots, \mathbf{p}_{i,m}]^\top \in \mathbb{R}^{m \times k}$. Thus, the local conditional parameters can be estimated by solving the log-likelihood function, defined as follows.

$$\begin{aligned}&\arg\max_{\mathbf{w}_i, \sigma_i^2} \log \mathbb{L}(\mathbf{w}_i, \sigma_i^2 : \mathbf{c}'_i, \mathbf{P}_i) = \\ &-\frac{1}{2}\sum_m \left[\log(2\pi\sigma_i^2) + \frac{c'_{i,m} - \mathbf{w}_i^\top \mathbf{p}_{i,m})^2}{\sigma_i^2}\right].\end{aligned} \tag{4}$$

By estimating the parameters for all local conditionals, we have the parameter of Bayesian network as $\theta_\mathcal{B} = [(\hat{\mathbf{w}}_1, \hat{\sigma}_1^2), \ldots, (\hat{\mathbf{w}}_n, \hat{\sigma}_n^2)]$ with $n = k$. The likelihood that the dependency structure formed by the observed variables and the latent code coincides the topology of $\mathcal{B}$ and is the joint product of local conditionals, calculated by:

$$\mathbb{P}(\mathbf{c}, \mathbf{c}'; \mathcal{B}) = \prod_i \mathbb{P}(c'_i \mid \mathbf{p}_i, \theta_\mathcal{B}) \mathbb{P}(c_i; \mathcal{N}(\mathbf{0}, 1)). \tag{5}$$

Notably, the likelihood in Eq. (5) takes the form of a factorization of local conditionals and thus measures how well $\mathcal{B}$ fits the generated samples. By maximizing Eq. (5), we intermediately manipulate the data generator, regularizing it to generate samples with respect to the fidelity and diversity principles as we desire. In particular, the synthesized samples

are faithful since 1) they cannot be differentiated by the discriminator and must belong to an existing class and 2) we can enforce the disentanglement of salient features by controlling which class they belong to, such that the minorities are synthesized precisely. Also, the synthesized samples can be diverse and carry various visual patterns because the salient features representing the visual concepts can be entangled in an informative, controllable, and hence interpretable fashion.

3.2 Gradient-Based Parameter Estimation

Estimating the parameter $\theta_{\mathcal{B}}$ for the Bayesian regularizer entails a series of estimations of local conditionals by solving Eq. (4). Unfortunately, this process could be tedious and may introduce an estimation *bias* depending on the number of minority samples in a training batch. It is expected to bypass this process and estimate the parameters in a stochastic way, so as to improve the efficiency and minimize the bias.

To this end, we inject the likelihood maximization process into the loop of GAN training, guiding the GAN to generate data with a desired structure among the extracted features. Our solution is built upon the following observation. Suggested by Eq. (2), a mapping from observed variables to the latent code is parameterized by θ_G and θ_F. By taking the first-order derivative of Eq. (4) *w.r.t.* $\mathbf{w}_i$ and σ_i^2, we observe that the estimation of $\theta_{\mathcal{B}}$ entails the marginal, along with the expectations between the observed and latent variables, indicating that θ_G and θ_F together with the factored distribution of the observed variables contain sufficient statistics to approximate $\theta_{\mathcal{B}}$. It allows to approximate the joint distribution $\mathbb{P}(\mathbf{c}, \mathbf{c}'; \mathcal{B})$ with θ_G and θ_F directly, omitting $\theta_{\mathcal{B}}$ estimation, where the likelihood function is expressed as $\mathbb{L}(\theta_G, \theta_F : \mathbf{c}, \mathbf{c}', \mathcal{B})$. Based on this observation and Eq. (5), we derive the objective function as follows.

$$\begin{aligned}
&\arg\max_{\theta_G, \theta_F} \log \mathbb{L}(\theta_G, \theta_F : \mathbf{c}, \mathbf{c}', \mathcal{B}) \\
&= \arg\min_{\theta_G, \theta_F} \sum_{i=1}^{|\mathbf{c}'|} \Bigg[\sum_{j=1}^{k} \ell\big(\mathbb{E}(c_i'), \mathbb{E}(\mathbf{p}_{i,j}); \theta_G, \theta_F\big) \\
&\qquad - \sum_{j=1}^{2|\mathbf{c}'|-k} \ell\big(\mathbb{E}(c_i'), \mathbb{E}(\bar{\mathbf{p}}_{i,j}); \theta_G, \theta_F\big) \Bigg], \qquad (6)
\end{aligned}$$

where $\mathbb{E}(c_i')$ and $\mathbb{E}(\mathbf{p}_{i,j})$ correspond to the empirical expectations of the i-th latent code and its k-th parent estimated in one training batch, respectively. Denoted by $\bar{\mathbf{p}}_i$ the values taken by the variables which are *not* the parents of c_i' in $\mathcal{B}$. The intuition behind Eq. (6) is as follows. The value of the minus of two loss terms, both evaluated by ℓ, shall decrease if the salient features are correlated with their parents only, *i.e.*, the parenting variables can precisely infer the latent code in the feedforward pass; Otherwise, if the salient features are correlated with their non-parents or cannot be inferred from their parents, this value increases. In practice, we implement the loss metric ℓ with mean squared error, which equates to the log-likelihood maximization for Gaussians in generalized linear models [McCullagh, 2018]. Thus far, Eq. (6) which functions as the Bayesian regularizer can be optimized through gradient-based, stochastic updates. By enforcing the observed variables and the latent codes to follow a specified structure (*i.e.*, $\mathcal{B}$), our GAN is guided to learn how to synthesize samples in an efficient and interpretable manner.

4 Experiments

This section experimentally validates that our approach can empower simple classifiers to perform the state-of-the-art imbalanced classification accuracy by synthesizing high-quality minority samples across various imbalance settings.

Datasets. We benchmark our experiments on two widely used image sets, namely, MNIST [LeCun *et al.*, 2010] and Fashion-MNIST [Xiao *et al.*, 2017]. To simulate an imbalance setting, we follow [Mullick *et al.*, 2019] to choose specific classes as the minority and then randomly remove samples from those classes until the imbalanced ratio (IR, ratio of the size of the largest class to that of the smallest class) reaches a preset threshold. To verify the robustness of our approach, we use 6 different IR settings ranging from 50 to 500. Notably, a dataset with IR over 10 can be deemed as *highly* imbalanced and over 100, as *extremely highly* imbalanced. The larger the IR, the more difficult the classification task. To the best of our knowledge, cost-sensitive learning methods thinly restrict the IR under 10, while oversampling methods usually allow higher IRs. We perform a 10-fold cross-validation to eliminate the randomization bias and record the averaged results and the corresponding statistics.

Compared Methods. We take three imbalanced learning competitors, namely, SMOTE [Chawla *et al.*, 2002], InfoGAN [Chen *et al.*, 2016], and DGC [Wang *et al.*, 2020]. Among them, SMOTE and InfoGAN represent the interpolation-based oversampling techniques via capturing linear relations and deep geometric relations among data, respectively. DGC denotes a state-of-the-art imbalanced classifier, which does not explicitly synthesize new samples but balances the sample counts of class priors in the latent feature space directly. Note, DGC operates in an end-to-end fashion. To conduct a fair comparison, we synthesize the minority samples using SMOTE, InfoGAN, and our approach, and then feed the balanced datasets (original plus synthesized samples) to three CNNs with an identical architecture for classification. Our evaluation aims to answer the following three research questions.

Q1. *Does our approach outperform the state-of-the-arts?*

Given that the minority classes are of greater interests in real-world applications, *recall*, which measures the percentage of minority samples being correctly classified, is used. Also, to test how well the methods can maintain an accurate prediction on the majority classes, *F1-score* that represents the harmonic mean of recall and precision is employed.

Table 1 presents the comparative results. We have four observations. *First*, our approach achieves the best performance with 87.3% recall and 91.69% F1-score on average. The statistical evidence exhibits that our approach outperforms its counterparts across 4 experimental settings, with 16% and 6.17% performance improvement on average in terms of recall and F1-score, respectively. *Second*, SMOTE performs the worst by achieving averagely 76.5% recall and 85.94%

Method	MNIST				Fashion-MNIST			
	8 (IR=67:1)		9 (IR=100:1)		Bag (IR=67:1)		Ankle boot (IR=100:1)	
	Recall	F1-Score	Recall	F1-Score	Recall	F1-Score	Recall	F1-Score
SMOTE	.829 ± .002 •	.899 ± .001 •	.767 ± .003 •	.860 ± .002 •	.686 ± .003 •	.812 ± .001 •	.778 ± .004 •	.866 ± .001 •
InfoGAN	.802 ± .003 •	.876 ± .000 •	.780 ± .003 •	.855 ± .001 •	.734 ± .003 •	.839 ± .001 •	.815 ± .001 •	.874 ± .000 •
DGC	.847 ± .001 •	.911 ± .000 •	.740 ± .001 •	.843 ± .000 •	.709 ± .001 •	.824 ± .000 •	.853 ± .000 •	.899 ± .000 •
Ours.	**.893 ± .001**	**.930 ± .000**	**.859 ± .000**	**.915 ± .000**	**.831 ± .001**	**.899 ± .000**	**.908 ± .001**	**.922 ± .000**

Table 1: Experimental results (Mean Accuracy ± Standard Deviation) with various IRs. The best results are bold. • indicates our approach significantly outperforms the compared method (statistically, with the hypothesis supported by *paired t-tests* at 95% significance level).

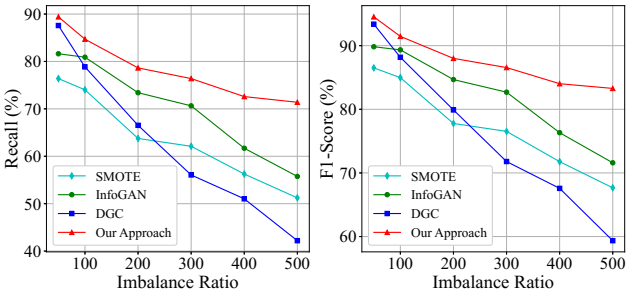


Figure 3: Classification results under different imbalance ratios, with metrics being **Left:** Recall and **Right:** F1-Score.

F1-score. This substantiates that the conventional oversampling techniques leveraging the linear structure to synthesize minority samples, cannot handle high-dimensional and complex data well, thereby yielding inferior performance.

Third, InfoGAN and DGC, both of which exploit deep models to capture the non-linear latent structure underlying data, enjoy better performances over SMOTE, with 2.3% and 2.92% recall improvement, respectively, and 0.27% and 1.2% F1-score improvement, respectively. This verifies the usefulness of modeling the data generating mechanism with salient features (latent code). *Fourth*, all methods perform worse in Fashion-MNIST than in MNIST, making an intuitive sense because Fashion-MNIST contains real-world data that carry more plentiful yet complex hidden patterns. We observe that InfoGAN and DGC decrease their recalls by 2.09% and 1.6% in Fashion-MNIST, respectively, from those in MNIST. This indicates that the data-hungriness of InfoGAN and DGC leads to inferior performance results in highly imbalanced settings, where the very limited sizes of minority classes cannot afford sufficient samples to allow them to properly characterize the complex patterns hidden behind Fashion-MNIST. On the contrary, our approach enjoys a higher sample efficiency by leveraging a Bayesian structure and hence performs robustly in Fashion-MNIST, with only a 0.76% drop in recall.

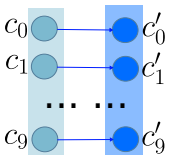
Q2. *How robustly can our approach perform in a variety of imbalance settings?*

We answer this question with Figure 3, in which the trends of recall and F1-score in different imbalance ratios are illustrated. Our observations are as follows. *First*, our approach outperforms the compared methods in terms of recall and F1-score across all IR settings. Also, as IR increases, the performance of our approach deteriorates less noticeably (for large IRs ≥ 400) than the compared methods. *Second*, DGC outperforms InfoGAN in low IR settings but its recall and F1-score drop more steeply than InfoGAN. Hence, InfoGAN exceeds DGC as IR increases. This is because DGC models a structural relationship between the latent code and the class labels in a closed form, missing out the intra-relationship among variables in the latent code. Therefore, DGC conceptually incurs high sample complexity, being more sensitive to different IRs. Both InfoGAN and our approach respect such variable-wise interactions and thus are more robust than DGC. *Third*, our approach enjoys a higher sample efficiency than InfoGAN, evidencing that our employed Bayesian structure can model the variable-wise interaction in a more explicit manner than mutual information.

Q3. *How interpretable are our extracted salient features, and how well are they under control?*

The answer to this question serves as an ablative study, revealing the effectiveness of our approach in implementing the *fidelity* and *diversity* principles during the minority synthesis process. To preserve the fidelity, we desire a strict disentanglement of the latent features deciding which classes the synthesized samples belong to. To this end, we apply a one-to-one dependency structure to the observed variables and the latent code, as shown in Figure 4a. Beyond a visual, qualitative validation, we quantify the disentanglement capacity of our approach by using the dSprites dataset, following the setups of [Kim and Mnih, 2017]. The disentanglement scores calculated for InfoGAN and our approach across various IRs are reported in Figure 4b. The higher the score, the better the disentangling performance, with our approach clearly outperforming InfoGAN in all IR settings.

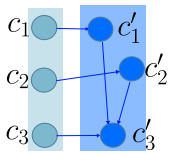
To verify whether our approach can synthesize samples with salient features being interacted in a pre-defined structure, we entangle the variables in the latent code as depicted in Figure 4c. The correspondingly synthesized samples are demonstrated in Figures 4d and 4e, where, the first two features extracted by the model in Figure 4d are thickness (c_1) and width (c_2), and the features in Figure 4e are rotation (c_1) and width (c_2). From the figures, we observe that the interaction of these features leads to an entangled latent variable c_3 which conveys the semantic meanings of c_1 and c_2 jointly. These results suggest that the Bayesian regularizer can guide GAN to entangle or disentangle extracted latent variables in accordance with a given dependency structure.



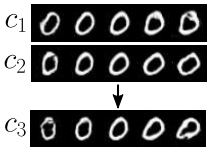
(a) Disentangled.

IR	100	200	300
InfoGAN	.525	.410	.361
Ours.	.585	.479	.388

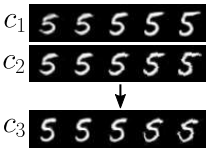
(b) dSprite scores.



(c) Entangled.



(d) Thick + Rotate.



(e) Size + Rotate.

Figure 4: Bayesian regularizer to control the salient feature space.

5 Related Work

As our work tightly relates to deep generative modeling and imbalanced classification, in this section, we discuss how it connects with and differs from the prior arts.

Deep Generative Models. Popularized by [Kingma and Welling, 2013] and [Goodfellow *et al.*, 2014], Variational Auto-Encoders (VAEs) and Generative Adversarial Networks (GANs) have become the *de facto* solutions to model the data generation mechanism with deep architectures. To compare, VAEs explicitly model the data density functions via variational approximations, while GANs implicitly learn the data densities by setting up a two-player game between a generator and a discriminator. Subsequent VAEs [Higgins *et al.*, 2017; Adel *et al.*, 2018] investigate how to further disentangle latent features by imposing constraints on the variational posterior. Compared to VAEs, GANs need to implement additional apparatus for the latent feature disentanglement. A prominent work is InfoGAN [Chen *et al.*, 2016], where the mutual information between the observed and the extracted latent feature subsets are maximized to realize the disentanglement. In addition to a set of subsequent studies such as [Tran *et al.*, 2017; Lee *et al.*, 2020], we are aware of two recent works [Kim *et al.*, 2019; Beyazit *et al.*, 2020] that also respect dependency structure among variables in the latent space with Bayesian treatment. However, these works were to extract various types of salient features while our task is imbalanced classification. The technical challenge and focus are thus different.

Despite effective, most existing deep generative models envision uniformly distributed data and usually perform poorly in an imbalanced setting with scarce minority samples [Ojha *et al.*, 2020]. Exploiting deep generative models for imbalanced classification is relatively under-explored. Our approach strives to fill this gap, enabling the extraction of interpretable and salient features with high sample efficiency, thereby enjoying a robust and quality minority synthesis across various imbalance settings.

Imbalanced Classification. Existing studies can be classified into two categories [Branco *et al.*, 2016; Huang *et al.*, 2016]. *First*, cost-sensitive learning, which tailors task-specific loss functions, such that the minority classes are more focused during the optimization process [Elkan, 2001]. Recent advances include focal loss [Lin *et al.*, 2017a] and dice loss [Li *et al.*, 2020a], which have manifested remarkable performance in the tasks of computer vision and natural language processing, respectively. Unfortunately, they entail extensive domain expert knowledge to craft the learning objectives and to tune the hyperparameters [Li *et al.*, 2020b], thereby being less accessible for users without such knowledge.

Second, resampling techniques, which manipulate the class priors by either dampening majorities (*i.e.*, undersampling) or synthesizing minorities (*i.e.*, oversampling) strategically. As undersampling tends to remove valuable information [Lin *et al.*, 2017b], especially in highly imbalanced settings, oversampling becomes superior.

However, the mainstream oversampling studies focus on generating artificial points with a linear combination of the existing data [Chawla *et al.*, 2002; He *et al.*, 2008; 2018; Yin *et al.*, 2020], failing to respect the fact that high-dimensional and complex data usually lie around a lower-dimensioned, non-linear manifold. Few studies have explored how to capture such intrinsic structures with deep models [Guo *et al.*, 2019; Wang *et al.*, 2020] with their methods operating in an end-to-end, black-box fashion, leaving the data generating mechanism agnostic and unknown. Therefore, these studies always suffer from tedious parameter-tuning processes based on various domains, tasks, and imbalance settings. Our approach opens the black box and explicitly models the data generating distribution in an interpretable feature space, controlled with human-in-the-loop.

6 Conclusion

This paper has proposed a novel oversampling approach which exploits the generative adversarial network (GAN) to model the data generating mechanism for the minority samples. The crux of our design lies in the learning of latent yet interpretable representations to capture the non-linear geometric structure underlying data. To this end, we have devised a Bayesian regularizer which guides the GAN to extract a set of salient features that interact in accordance with a dependency structure given a priori. As such, we can control over the learned space by intentionally entangle or disentangle the salient feature as we wish, so as to synthesize minority samples that carry diverse patterns and are faithful to the class labels. We have carried out both quantitative and qualitative experiments in a variety of imbalanced settings. The results substantiate that 1) our approach can synthesize high-quality samples in even extremely high imbalance ratios, and 2) our synthesized samples can help the simple classifiers to outperform the state-of-the-art imbalanced learning methods.

Acknowledgements

The authors would like to thank the IJCAI 2021 reviewers for their constructive feedback. This work was supported in part by the US National Science Foundation (NSF) under Grants 1652107, 1763620, 1948374, and 2019511, and in part by the Louisiana Board of Regents under Contract LEQSF(2018-21)-RD-A-24. Any opinion and findings expressed in the paper are those of the authors and do not necessarily reflect the view of funding agencies.

References

[Adel *et al.*, 2018] Tameem Adel, Zoubin Ghahramani, and Adrian Weller. Discovering interpretable representations for both deep generative and discriminative models. In *ICML*, pages 50–59, 2018.

[Beyazit *et al.*, 2020] Ege Beyazit, Doruk Tuncel, Xu Yuan, Nian-Feng Tzeng, and Xindong Wu. Learning interpretable representations with informative entanglements. In *IJCAI*, pages 1970–1976, 2020.

[Branco *et al.*, 2016] Paula Branco, Luís Torgo, and Rita P Ribeiro. A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys (CSUR)*, 49(2):1–50, 2016.

[Chawla *et al.*, 2002] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[Chen *et al.*, 2016] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *NeurIPS*, 29:2172–2180, 2016.

[Elkan, 2001] Charles Elkan. The foundations of cost-sensitive learning. In *IJCAI*, volume 17, pages 973–978. Lawrence Erlbaum Associates Ltd, 2001.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NeurIPS*, 27:2672–2680, 2014.

[Guo *et al.*, 2019] Ting Guo, Xingquan Zhu, Yang Wang, and Fang Chen. Discriminative sample generation for deep imbalanced learning. In *IJCAI*, pages 2406–2412, 2019.

[He *et al.*, 2008] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *IJCNN*, pages 1322–1328. IEEE, 2008.

[He *et al.*, 2018] Yi He, Di Wu, Ege Beyazit, Xiaoduan Sun, and Xindong Wu. Supervised data synthesizing and evolving–a framework for real-world traffic crash severity classification. In *ICTAI*, pages 163–170. IEEE, 2018.

[Heckerman and Wellman, 1995] David Heckerman and Michael P Wellman. Bayesian networks. *Communications of the ACM*, 38(3):27–31, 1995.

[Higgins *et al.*, 2017] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

[Huang *et al.*, 2016] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *CVPR*, pages 5375–5384, 2016.

[Kim and Mnih, 2017] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *NeurIPS worshop: From Perception to Control*, 2017.

[Kim *et al.*, 2019] Minyoung Kim, Yuting Wang, Pritish Sahu, and Vladimir Pavlovic. Bayes-factor-vae: Hierarchical bayesian deep auto-encoder models for factor disentanglement. In *CVPR*, pages 2979–2987, 2019.

[Kingma and Welling, 2013] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[LeCun *et al.*, 2010] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. 2010.

[Lee *et al.*, 2020] Wonkwang Lee, Donggyun Kim, Seunghoon Hong, and Honglak Lee. High-fidelity synthesis with disentangled representation. In *ECCV*, pages 157–174. Springer, 2020.

[Li *et al.*, 2020a] Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. Dice loss for data-imbalanced nlp tasks. In *ACL*, pages 465—476, 2020.

[Li *et al.*, 2020b] Yu Li, Tao Wang, Bingyi Kang, Sheng Tang, Chunfeng Wang, Jintao Li, and Jiashi Feng. Overcoming classifier imbalance for long-tail object detection with balanced group softmax. In *CVPR*, pages 10991–11000, 2020.

[Lin *et al.*, 2017a] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017.

[Lin *et al.*, 2017b] Wei-Chao Lin, Chih-Fong Tsai, Ya-Han Hu, and Jing-Shang Jhang. Clustering-based undersampling in class-imbalanced data. *Information Sciences*, 409:17–26, 2017.

[McCullagh, 2018] Peter McCullagh. *Generalized linear models*. Routledge, 2018.

[Mullick *et al.*, 2019] Sankha Subhra Mullick, Shounak Datta, and Swagatam Das. Generative adversarial minority oversampling. In *CVPR*, pages 1695–1704, 2019.

[Ojha *et al.*, 2020] Utkarsh Ojha, Krishna Kumar Singh, Cho-Jui Hsieh, and Yong Jae Lee. Elastic-infogan: Unsupervised disentangled representation learning in class-imbalanced data. *NeurIPS*, 2020.

[Tran *et al.*, 2017] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*, pages 1415–1424, 2017.

[Wang *et al.*, 2020] Xinyue Wang, Yilin Lyu, and Liping Jing. Deep generative model for robust imbalance classification. In *CVPR*, pages 14124–14133, 2020.

[Xiao *et al.*, 2017] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[Yin *et al.*, 2020] Jian Yin, Chunjing Gan, Kaiqi Zhao, Xuan Lin, Zhe Quan, and Zhi-Jie Wang. A novel model for imbalanced data classification. In *AAAI*, pages 6680–6687, 2020.