# Generative Oversampling for Imbalanced Data via Majority-Guided VAE

**Qingzhong Ai**[1] **Pengyun Wang**[2,*] **Lirong He**[1] **Liangjian Wen**[2]
**Lujia Pan**[2] **Zenglin Xu**[3,4,†]

[1]University of Electronic Science and Technology of China [2]Noah's Ark Lab, Huawei Technologies
[3]Harbin Institute of Technology, Shenzhen [4]Peng Cheng Lab

{qzai,lirong_he}@std.uestc.edu.cn wangpyun1203@gmail.com
{wenliangjian1,panlujia}@huawei.com xuzenglin@hit.edu.cn

## Abstract

Learning with imbalanced data is a challenging problem in deep learning. Over-sampling is a widely used technique to re-balance the sampling distribution of training data. However, most existing over-sampling methods only use intra-class information of minority classes to augment the data but ignore the inter-class relationships with the majority ones, which is prone to overfitting, especially when the imbalance ratio is large. To address this issue, we propose a novel over-sampling model, called Majority-Guided VAE (MGVAE), which generates new minority samples under the guidance of a majority-based prior. In this way, the newly generated minority samples can inherit the diversity and richness of the majority ones, thus mitigating overfitting in downstream tasks. Furthermore, to prevent model collapse under limited data, we first pre-train MGVAE on sufficient majority samples and then fine-tune based on minority samples with Elastic Weight Consolidation (EWC) regularization. Experimental results on benchmark image datasets and real-world tabular data show that MGVAE achieves competitive improvements over other over-sampling methods in downstream classification tasks, demonstrating the effectiveness of our method.

## 1 Introduction

Modern advanced models, such as deep neural networks (DNNs), are driven by large-scale training data of high quality, which is usually well-designed and class-balanced. However, the distributions of real-world data tend to be more complex. For example, some classes of a dataset are difficult to access due to scarcity or privacy, resulting in a significant difference in the number of training instances from the other classes. This is typically known as the class-imbalanced (Johnson and Khoshgoftaar, 2019; Huang et al., 2019; Wang et al., 2022) or "long-tailed" (Mahajan et al., 2018; Van Horn et al., 2018; Zhang et al., 2021) problem. Generally, we refer to the class with sufficient samples as the majority class, and the class with few data as the minority class. On these class-imbalanced datasets, the standard training of DNNs has been found to perform poorly (Wang et al., 2017; Dong et al., 2018; Ren et al., 2018; He and Garcia, 2009), especially in classification tasks. The resultant classification surface of the classifier tends to be highly skewed towards the majority class due to its dominance (Das et al., 2018). As a result, the prediction accuracy in the minority classes is drastically affected, and the overall generalization performance suffers.

To alleviate the detrimental effects of the class-imbalanced problem, one typically needs to re-balance the training objective with respect to class-wise instance size via two basic approaches: re-weighting and re-sampling. The principal concept of the re-weighting (Khan et al., 2017; Chung et al., 2015; Lin et al., 2017) is to adjust the objective learning function so that the samples in the minority classes receive more attention than the majority ones. Hence, it is also called the objective-level method. Many advanced re-weighting techniques have been proposed, such as RW (Huang et al., 2016), CBRW (Cui et al., 2019), FOCAL (Lin et al., 2017), and LDAM Cao et al. (2019) etc. Unlike re-weighting, re-sampling focuses on obtaining a balanced sampling distribution during training, which could be achieved by either "under-sampling" the majority

[*] Work done while at Huawei.
[†] Corresponding Author.

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206. Copyright 2023 by the author(s).

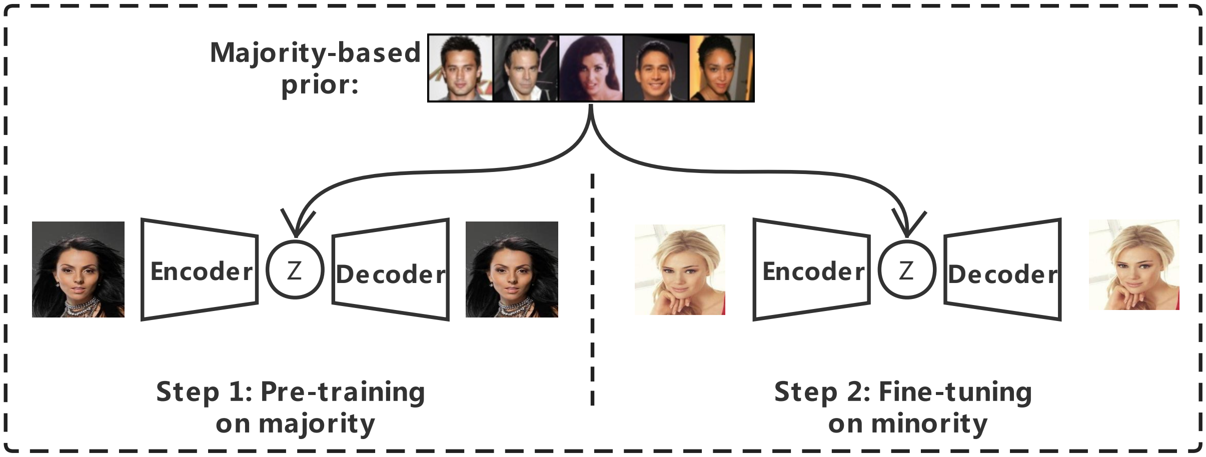

Majority-based prior:
Encoder
Z
Decoder
Step 1: Pre-training on majority
Encoder
Z
Decoder
Step 2: Fine-tuning on minority

Figure 1: Pipeline of our method. Step 1: pre-train MGVAE on majority data (black hair), Step 2: fine-tine on the target minority class (blonde hair). Both steps under the majority-based prior.

classes (He and Garcia, 2009; Liu et al., 2008) or "over-sampling" the minority classes (Cui et al., 2018; Japkowicz, 2000; Kang et al., 2019; Wang et al., 2020a). So, it belongs to the data-level method.

In this paper, we focus on the technique of "over-sampling." The simplest way is Random Over Sampling (ROS) (Japkowicz, 2000). As a representative algorithm, SMOTE (Chawla et al., 2002) and its variants (Han et al., 2005; He et al., 2008; Mullick et al., 2019) augment the minority class through intra-class linear interpolation sampling to alleviate overfitting, which has been widely used. However, when the number of samples in the minority class reduces to a few, the performance of SMOTE drops noticeably. Despite the recent influx of advanced over-sampling methods (Fajardo et al., 2021; Zhang and Pfister, 2021; Wang et al., 2020b), we notice that most of them only use intra-class information of minority classes to augment the data while ignoring the inter-class relationships, which invites effort to make further improvements. Intuitively, the majority and minority classes of the same dataset can be related, and therefore the latter can borrow the diversity and richness of the former for the purpose of a more informative data augmentation.

To this end, we propose Majority-Guided VAE (MGVAE), a novel over-sampling model for generating the minority samples under the guidance of a majority-based prior. Specifically, we first model the distribution of the minority as a parametric distribution conditioned on the majority samples. Then, by introducing a parametric transition distribution with a hidden variable, the learning objective is tractable through variational inference, resulting in a new VAE framework. In essence, MGVAE is a generative model for minority classes with a majority-based prior. Nevertheless, minority samples cannot satisfy the requirement of large-scale data for the training of deep generative models to guarantee generation quality (Kingma and Welling, 2013; Goodfellow et al., 2014; Dinh et al., 2014; Ho et al., 2020; Ai et al., 2021; Fan et al., 2022). Limited data can lead to overfitting or even model collapse. To fix this flaw, we take inspiration from the paradigm of Few Shot Generation (FSG) (Li et al., 2020; Ojha et al., 2021; Wang et al., 2018) and adopt a pre-training and fine-tuning framework, as shown in Figure 1. In detail, we first pre-train MGVAE on the majority's samples, which are sufficient for training a VAE from sketch. Then, we move on to fine-tune the pre-trained model based on the minority data. In addition, to prevent catastrophic forgetting of the majority and overfitting to the minority, we adopt the Elastic Weight Consolidation (EWC) to regulate the fine-tuning progress, which proves to be beneficial. The resultant model can be used to generate new minority data by: 1) first drawing points at random from the majority based prior in the latent space; 2) and then transforming them through the learned decoder to obtain new minority data.

As we will see, the over-sampling of minority classes is straightforward with a trained MGVAE. It is able to perform one-to-one "translation" of the instances from majority to minority, resulting in a class-balanced training dataset. To evaluate the over-sampling quality of MGVAE, we conduct extensive experiments based on various classifier backbones and imbalanced datasets. Compared to previous over-sampling techniques, our model performs the best in all cases.

## 2 Methods

In the big picture, we consider a classification problem on a class-imbalanced dataset $\mathcal{D}_{imb}$. Note that our method is orthogonal to the downstream tasks and thus can be seamlessly integrated into any downstream tasks without multiple training. For a clear elaboration, we will introduce our method by confining attention to a binary dataset with a majority class $\mathcal{X}^+ \equiv \{\mathbf{x}_n^+\}_{n=1}^{N^+}$ and a minority class $\mathcal{X}^- \equiv \{\mathbf{x}_n^-\}_{n=1}^{N^-}$. Generalizing to multi-class is straightforward and will be discussed later. In the setting of class-imbalanced, we have $N^+ \gg N^-$, where $N^+$ and $N^-$ denote the sample size of the majority and minority. Throughout this paper, vectors are symbolized by bold lowercase letters whose subscripts indicate their order, and matrices are denoted by upper-case letters.

### 2.1 Majority-Guided VAE

In general, sample size correlates with diversity. Compared to the minority class, the majority has a large-scale sample size, which means better diversity and richer information. We argue that it is crucial to be able to borrow the knowledge of the majority when augmenting the minority. To this end, we formulate the distribution of the minority as a parametric distribution conditioned on the majority. Formally, the log density distribution can be expressed as

$$\log p(\mathbf{x}^- \mid \mathcal{X}^+, \Psi), \tag{1}$$

where $\mathbf{x}^-$ denotes the minority sample, $\mathcal{X}^+ \equiv \{\mathbf{x}_n^+\}_{n=1}^{N^+}$ is the set of majority samples, and $\Psi$ is the distribution parameter. More specifically, we define a parametric transition distribution $T_\Psi(\mathbf{x}^- \mid \mathbf{x}^+)$, which stochastically transforms a majority sample $\mathbf{x}^+$ into a new minority observation $\mathbf{x}^-$. Then, we have

$$\log p(\mathbf{x}^- \mid \mathcal{X}^+, \Psi) = \log \sum_{n=1}^{N^+} \frac{1}{N^+} T_\Psi(\mathbf{x}^- \mid \mathbf{x}_n^+), \tag{2}$$

where the prior probability is uniform over the majority sample. Choosing a suitable transition distribution is crucial for better approximating the data distribution, especially for a small number of observations. For example, the Kernel Density Estimator (KDE) can be seen as a simple non-parametric transition distribution, but with limited expressive power. For the pursuit of powerful expressive ability, we introduce a parameter transition distribution with the latent variable $\mathbf{z}$, which defined as

$$T_\Psi(\mathbf{x}^- \mid \mathbf{x}^+) = \int_z r_\phi(\mathbf{z} \mid \mathbf{x}^+) p_\theta(\mathbf{x}^- \mid \mathbf{z}) \mathbf{dz}, \tag{3}$$

where the parameters $\Psi = \{\theta, \phi\}$ and $r_\phi(\mathbf{z} \mid \mathbf{x}^+)$ is a prior based on majority. We assume that a minority observation $\mathbf{x}^-$ is independent to a majority sample $\mathbf{x}^+$ conditional on $\mathbf{z}$ for simplifying the formulation and optimization. In fact, a generative model of the minority is already embedded in the transition distribution in Eq. (3). $r_\phi(\mathbf{z} \mid \mathbf{x}^+)$ is a majority-based prior with parameter $\phi$ for generating latent code $\mathbf{z}$ from a majority sample $\mathbf{x}^+$. Then, a decoder $p_\theta(\mathbf{x}^- \mid \mathbf{z})$ with parameter $\theta$ generates new minority sample $\mathbf{x}^-$ from $\mathbf{z}$. Next, we can optimize the evidence lower bound (ELBO) derived by variational inference as follows,

$$\begin{aligned}
&\log p(\mathbf{x}^- \mid \mathcal{X}^+, \Psi) \\
&= \log \sum_{n=1}^{N^+} \frac{1}{N^+} T_\Psi(\mathbf{x}^- \mid \mathbf{x}_n^+) \\
&= \log \sum_{n=1}^{N^+} \frac{1}{N^+} \int_z r_\phi(\mathbf{z} \mid \mathbf{x}_n^+) p_\theta(\mathbf{x}^- \mid \mathbf{z}) \mathbf{dz} \\
&\overset{\text{VAE}}{\geq} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}^-)} \log p_\theta(\mathbf{x}^- \mid \mathbf{z}) \\
&- \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}^-)} \log \frac{q_\phi(\mathbf{z} \mid \mathbf{x}^-)}{\sum_{n=1}^{N^+} r_\phi(\mathbf{z} \mid \mathbf{x}_n^+)/N^+} \\
&\equiv O(\Psi, \mathcal{X}^+; \mathbf{x}^-),
\end{aligned} \tag{4}$$

where $O(\Psi, \mathcal{X}^+; \mathbf{x}^-)$ is the optimization objective of our proposed model, MGVAE. See detailed derivation in the supplementary materials A. According to Eq. (4), we know that MGVAE is a generative model with a VAE-like framework for the minority, where the approximate posterior $q_\phi(\mathbf{z} \mid \mathbf{x}^-)$ is the encoder and $p_\theta(\mathbf{x}^- \mid \mathbf{z})$ is the decoder. Similar to the standard VAE, the first part of the objective is the reconstruction error. The difference lies in the KL divergence of the second part, where the prior of MGVAE is a mixture prior $p(\mathbf{z} \mid \mathcal{X}^+) = \sum_{n=1}^{N^+} r_\phi(\mathbf{z} \mid \mathbf{x}_n^+)/N^+$, in which each component is conditioned on a majority sample. To generate a new observation from the pre-trained MGVAE, we need both the decoder network $p_\theta$ and the prior network $r_\phi$. The generating process of MGVAE is summarised in Algorithm 1. As shown, MGVAE can achieve a one-to-one probabilistic mapping from majority to minority, resulting in a class-balanced dataset.

**Algorithm 1:** The Generating Process of MGVAE.

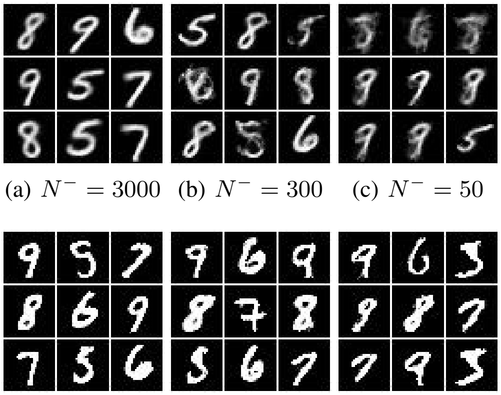
```
Input  : The majority samples $\mathcal{X}^+$, the decoder $p_\theta$,
         and the prior $r_\phi$.
Output : A generated observation $\mathbf{x}^-$.
Step.1 Sample $n \sim$ Unifrom$(0, N-1)$ for obtaining a
  random sample $\mathbf{x}_n^+$ from the majority $\mathcal{X}^+$.
Step.2 Sample $\mathbf{z} \sim r_\phi(\cdot \mid \mathbf{x}_n^+)$ using the majority point
  based prior $r_\phi$ to obtain a latent code $\mathbf{z}$.
Step.3 Sample $\mathbf{x}^- \sim p_\theta(\cdot \mid \mathbf{z})$ using the decoder $p_\theta$ for
  generating a new observation $\mathbf{x}^-$.
```

**Implemental Details** In practice, following advanced works, such as VampPrior, Exemplar VAE, and ByPE-VAE, the prior $r_\phi$ is modeled as a Gaussian distribution $\mathcal{N}(\mathbf{z} \mid \mu_\phi(\mathbf{x}), \sigma^2 I)$, in which the parametric mean function $\mu_\phi$ is shared with encoder $q$, and the covariance function is an isotropic Gaussian with a scalar parameter $\sigma$. Therefore, the parameters in $r_\phi$ can be updated every iteration by the backpropagation gradient descent. Moreover, instead of using entire majority samples, we randomly down-sample a fix-sized majority in each step to compute the prior for computation efficiency. The detailed training process is summarized in Algorithm 2 in supplementary materials C.

**Generalization to multi-class** It is straightforward to generalize MGVAE to the multiple class-imbalanced classification problems, such as long-tailed learning. For a long-tailed dataset $\mathcal{D}_{LT}$ with $K$ classes $\{C_i\}_{i=1}^{K}$, we have $N_1 > N_2 > ... > N_K$, where $N_i$ is the sample size of class $C_i$. To over-sampling the dataset $\mathcal{D}_{LT}$ with MGVAE, we construct $K-1$ binary class datasets. The majority class of each binary dataset is $C_1$, while the minority is $C_2 - C_K$, respectively. After that, we can obtain $K-1$ MGVAEs, each corresponding to one minority class. Finally, the over-sampling of the dataset $\mathcal{D}_{LT}$ is completed by randomly sampling each MGVAE according to Algorithm 1, resulting in a balanced dataset.

### 2.2 Training with limited data

So far, we have obtained a generative model and corresponding generating process for the minority with a majority-based prior. Unfortunately, the model does not work well yet to generate samples with limited data. We test MGVAE on modified-MNIST, where all 0-4 are used as the majority and downsampled 5-9 as the minority. The sampling results are shown in Figure 2(a) - 2(c), from which we can find that as the size of minority samples decreases, generation quality gradually deteriorates until indistinguishable. We argue this is a generative dilemma for all modern deep generative models. When the number of observations is much smaller than model parameters, the model tends to overfit or collapse.



(a) $N^- = 3000$ (b) $N^- = 300$ (c) $N^- = 50$

(d) $N^- = 300$ (e) $N^- = 50$ (f) $N^- = 10$

Figure 2: The sampling results of MGVAE. (a-c): Without the Pre-training and Fine-tuning; (d-f): With the Pre-training and Fine-tuning. The label under each figure is the size of the downsampled minority samples.

#### 2.2.1 Pre-training and Fine-tuning

To address this issue, we adopt the process of pre-training and fine-tuning, which is borrowed from the idea of FSG. Specifically, we first pre-train an MGVAE based on the majority samples, which is adequate for the training from sketch. After convergence, we move on to the minority, fine-tuning the pre-trained model based on limited minority samples. The pipeline of the training process is shown in Figure 1. Likewise, we can sample the model after the training process of "pre-training and fine-tuning" according to Algorithm 1. The sampling results are shown in Figure 2(d) - 2(f). By comparing with Figure 2(a) - 2(c), we find that the samples generated by the current model are meaningful and distinguishable, even when the size of the minority sample is extremely small (e.g., 10). More visual results are presented in the experimental section.

#### 2.2.2 EWC regularization

Further, we employ Elastic Weight Consolidation (EWC) regularization during the fine-tuning process. EWC is a commonly used fine-tune technique, first proposed in Kirkpatrick et al. (2017) to avoid catastrophic forgetting in continuous learning, and Li et al. (2020) introduces EWC to the FSG setting. In Nguyen et al. (2017), EWC is introduced for VAE in Variational Continual Learning. To prevent overfitting to the new domain and the catastrophic forgetting of the old one, EWC preserves the important parameters by penalizing parameter changes. The importance of each parameter is measured by Fisher Information, defined as follows,

$$F = \mathbb{E}\left[-\frac{\partial^2}{\partial \Psi^2}\mathcal{L}\left(X \mid \Psi\right)\right], \tag{5}$$

where $\mathcal{L}$ is the log-likelihood function that can be approximated by ELBO in VAE, $X$ is a collection of generated data, and $\Psi$ is the learned parameter value. Given the model pre-trained in the majority with learned parameters $\Psi_+$, we could get the Fisher Information vector $F$ and then formalize the fine-tuning process as

$$O_{\text{EWC}} = O(\Psi, \mathcal{X}^+; \mathbf{x}^-) + \lambda \sum_i F_i \left(\Psi_i - \Psi_{+,i}\right)^2, \tag{6}$$

where $O(\Psi, \mathcal{X}^+; \mathbf{x}^-)$ is the optimization objective in Eq. (4), $\lambda$ is the regularization weight, and $i$ is the index of each model parameter. Note that $O_{\text{EWC}}$ is our final optimization objective function.

## 3 Experiments

To evaluate the over-sampling quality of MGVAE, we consider classification as our downstream evaluation task. Given an imbalanced dataset $\mathcal{D}_{imb}$, we first augment the dataset by MGVAE to obtain a balanced dataset $\mathcal{D}_{bal}$. Then, we train the classifier based on the augmented dataset $\mathcal{D}_{bal}$. We test our model on several datasets of different scales and data types, including image datasets MNIST (LeCun et al., 1998), FashionMNIST (Xiao et al., 2017), CelebA (Liu et al., 2015), and several tabular datasets, namely Musk (Asuncion and Newman, 2007), Water Quality, and Isolet. We use different network architectures of MGVAE depending on the dimensionality and size of the dataset. The same goes for the classifier. For the image data, we also give the corresponding visual results, aiming to better demonstrate the significance and validity of our method. Finally, some analysis experiments, including ablation study and sensitivity analysis, are conducted to understand the proposed method in detail.

**Baseline Methods.** We adopt a wide range of existing methods as our baselines, as follows:

1. **Empirical risk minimization (ERM)**: the standard training method with cross-entropy loss and without any balancing operation.
2. **Random over-sampling (ROS)** (Japkowicz, 2000): balancing the sampling distribution by repeated random sampling of the minority class.
3. **SMOTE** (Chawla et al., 2002): balancing the sampling distribution by linear interpolating nearest neighbors in the minority class.
4. **Re-weighting (RW)** (Huang et al., 2016): modifying the objective function according to the class sample size.
5. **Class-balanced re-weighting (CBRW)** (Cui et al., 2019): an improved variant of RW, introducing the effective sample number $E_k = (1 - \beta^{N_k})/(1 - \beta)$ for each class, where $\beta$ is set to 0.9999.
6. **FOCAL** (Lin et al., 2017): aiming to balance the sample-wise classification loss for model training by down-weighing the well-classified samples.
7. **LDAM** (Cao et al., 2019): a label-distribution-aware margin loss that encourages few-shot classes to have larger margins.
8. **OCVAE** (Fajardo et al., 2021): **O**ver-sampling the minority by a **C**onditional **VAE**, a generative model augmenting the dataset conditioned on the class label.
9. **OCGAN** or **OCDCGAN** (Fajardo et al., 2021): **O**ver-sampling the minority by **C**onditional **GAN**, another generative model augmenting the dataset conditioned on the class label.

Note that the main network architecture of OCVAE, OCGAN, and OCDCGAN is consistent with our model for fairness. To distinguish all comparison methods more clearly, we classify them into four categories, classified as traditional over-sampling methods (ROS, SMOTE); re-weighting methods (RW, CBRW); other loss functions (FOCAL, LDAM), and DGMs-based over-sampling methods (OCVAE, OCGAN, and OCDCGAN).

**Datasets Information.** We test our model on both benchmark image datasets and real-world tabular data. For the image dataset, we chose MNIST, FashionMNIST, and CelebA, all of which are artificial class-balanced datasets. Therefore, we need to downsample the dataset to satisfy the class-imbalance setting. Similar to Fajardo et al. (2021), we conduct a challenging binary classification task for MNIST and FashionMNIST. Taking MNIST as an example, we take all classes of 0-4 as the majority classes with size 30000=6000*5, and downsample classes of 5-9 to 300=60*5 and 50=10*5 as the minority class. i.e., the imbalance ratio (IR) is $\rho = 100$ and $\rho = 600$, respectively. The processing of FashionMNIST is similar. For CelebA, we chose hair color as the label to form a five-class long-tail dataset, where the sizes of black hair, blonde hair, blad hair, brown hair, and gray hair are 20000, 10000, 2500, 1000, and 200, respectively. Black hair is used as the majority class. In addition, we also conducted experiments on several real-world tabular datasets, which have natural class imbalance and therefore do not require additional manipulations. Information of all datasets is summarised in Table 1.

Table 1: Data Information. “Maj” stands for majority, “Min” stands for minority, and “c” stands for class. The number X in “*Dataset*-X” stands for imbalance ratio.

| Datasets | | Split Way |
|---|---|---|
| Images | MNIST-100 | Maj: 0-4 (30000); Min: 5-9 (300) |
| | MNIST-600 | Maj: 0-4 (30000); Min: 5-9 (50) |
| | FashionMNIST-100 | Maj: c0-c4 (30000); Min: c5-c9 (300) |
| | FashionMNIST-600 | Maj: c0-c4 (30000); Min: c5-c9 (50) |
| | CelebA-100 | Long tail: 20000:10000:2500:1000:200 |
| Tabular | Musk-6.6 | Maj:c0 (5381); Min: c1 (817) |
| | Water Quality-11.1 | Maj: c0 (6784); Min: c1 (612) |
| | Isolet-17.5 | Maj: c0 (6997); Min: c1 (400) |

**Evaluation Metrics.** We use three evaluation metrics to assess the classification performance of all methods under the same balanced test distribution, namely Balanced Accuracy (B-ACC) (Huang et al., 2016; Wang et al., 2017; Kim et al., 2020), Average Class Specific Accuracy (ACSA) (Huang et al., 2016; Wang et al., 2017; Mullick et al., 2019), and Geometric Mean (GM) (Kubat et al., 1997; Branco et al., 2016). B-ACC is same as the standard accuracy metric on the balanced dataset. The other two metrics are not biased toward any particular class, therefore more suitable for evaluating the performance under an imbalanced setting. Note that all numerical results in this section average three random trials.

## 3.1 Results

We evaluate the performance of all methods in the following two aspects: the numerical results on classification as a quantitative comparison, and the other is the visual qualitative results. All results tables highlight the best results in bold, and the second-best results are underlined.

### 3.1.1 Quantitative Results

**MNIST and FashionMNIST.** Due to the similarity of the dataset scale (28*28), the experimental setups of MNIST and FashionMNIST are basically the same. We adopt two imbalance ratios, $\rho = 100$ and $\rho = 600$, for each dataset. The classifier is a 2-layer fully-connected neural network with 256 and 128 middle nodes, trained for 100 epochs with mini-batch size 100. The learning rate is initialized to $1e$-3 and decreases with an exponential schedule of $\gamma = 0.95$. Besides, the architecture of all generative models, including MGVAE and OCVAE, remains consistent to ensure fairness. The EWC regularization weight $\lambda$ in MGVAE is selected from the set of candidate $\{5e2, 5e4, 5e6, 5e8\}$. Notice that we omit the results of OCGAN due to the inability to train properly on small-scale data.

Table 2: Comparison of classification performance on MNIST with two imbalance radios.

| IR | $\rho = 100$ | | | $\rho = 600$ | | |
|---|---|---|---|---|---|---|
| Methods | B-ACC | ASCA | GM | B-ACC | ASCA | GM |
| ERM | $51.4 \pm 0.0$ | $50.0 \pm 0.0$ | $0.0 \pm 0.0$ | $51.4 \pm 0.0$ | $50.0 \pm 0.0$ | $0.0 \pm 0.0$ |
| FOCAL | $51.4 \pm 0.0$ | $50.0 \pm 0.0$ | $0.0 \pm 0.0$ | $51.4 \pm 0.0$ | $50.0 \pm 0.0$ | $0.0 \pm 0.0$ |
| RW | $77.4 \pm 1.2$ | $76.7 \pm 1.2$ | $73.1 \pm 1.3$ | $59.9 \pm 1.6$ | $58.8 \pm 1.4$ | $41.7 \pm 1.8$ |
| CBRW | $75.1 \pm 0.8$ | $74.3 \pm 0.7$ | $69.8 \pm 1.3$ | $56.1 \pm 0.5$ | $55.1 \pm 0.3$ | $31.2 \pm 1.2$ |
| LDAM | $82.9 \pm 0.5$ | $82.4 \pm 0.6$ | $80.3 \pm 0.7$ | $63.1 \pm 0.9$ | $62.0 \pm 0.8$ | $48.7 \pm 1.0$ |
| RS | $79.2 \pm 0.3$ | $78.5 \pm 0.3$ | $75.4 \pm 0.2$ | $58.5 \pm 1.0$ | $57.3 \pm 1.1$ | $37.7 \pm 2.0$ |
| SMOTE | $80.6 \pm 0.3$ | $80.1 \pm 0.2$ | $77.4 \pm 0.1$ | $60.0 \pm 0.9$ | $58.3 \pm 1.0$ | $40.2 \pm 1.1$ |
| OCVAE | $83.0 \pm 0.4$ | $82.6 \pm 0.4$ | $80.6 \pm 0.6$ | $63.8 \pm 0.2$ | $62.8 \pm 0.5$ | $50.7 \pm 0.5$ |
| **MGVAE** | **85.0** $\pm 0.2$ | **84.6** $\pm 0.2$ | **83.2** $\pm 0.2$ | **65.4** $\pm 1.0$ | **64.4** $\pm 1.1$ | **53.4** $\pm 1.1$ |

Table 3: Comparison of classification performance on FashionMNIST with two imbalance radios.

| IR | $\rho = 100$ | | | $\rho = 600$ | | |
|---|---|---|---|---|---|---|
| Methods | B-ACC | ASCA | GM | B-ACC | ASCA | GM |
| ERM | $86.4 \pm 0.1$ | $86.4 \pm 0.1$ | $85.3 \pm 0.1$ | $80.7 \pm 0.3$ | $80.8 \pm 0.4$ | $78.3 \pm 0.4$ |
| FOCAL | $86.9 \pm 0.4$ | $86.9 \pm 0.4$ | $85.9 \pm 0.4$ | $81.9 \pm 0.4$ | $82.0 \pm 0.3$ | $79.9 \pm 0.4$ |
| RW | $87.8 \pm 0.5$ | $87.8 \pm 0.5$ | $86.9 \pm 0.6$ | $83.6 \pm 0.9$ | $83.6 \pm 0.9$ | $81.9 \pm 1.1$ |
| CBRW | $86.9 \pm 0.2$ | $87.0 \pm 0.3$ | $85.9 \pm 0.3$ | $82.8 \pm 0.9$ | $82.8 \pm 0.9$ | $81.0 \pm 1.1$ |
| LDAM | $87.7 \pm 0.2$ | $87.6 \pm 0.3$ | $86.8 \pm 0.3$ | $83.7 \pm 0.3$ | $83.8 \pm 0.3$ | $81.9 \pm 0.2$ |
| RS | $87.6 \pm 0.1$ | $87.5 \pm 0.2$ | $86.6 \pm 0.2$ | $81.7 \pm 0.7$ | $81.7 \pm 0.8$ | $79.5 \pm 0.5$ |
| SMOTE | $86.4 \pm 0.3$ | $86.5 \pm 0.3$ | $85.3 \pm 0.4$ | $80.7 \pm 0.8$ | $81.1 \pm 0.5$ | $78.7 \pm 0.6$ |
| OCVAE | $86.7 \pm 0.3$ | $86.7 \pm 0.3$ | $85.8 \pm 0.4$ | $82.8 \pm 0.4$ | $82.8 \pm 0.4$ | $81.1 \pm 0.5$ |
| **MGVAE** | **88.3** $\pm 0.1$ | **88.4** $\pm 0.1$ | **87.6** $\pm 0.1$ | **84.8** $\pm 0.4$ | **84.8** $\pm 0.4$ | **83.6** $\pm 0.4$ |

The main results for MNIST and FashionMNIST are presented in Table 2 and Table 3, respectively. Overall, the results show that our method consistently leads by a large margin compared to other tested methods. In particular, our method performs better than all other over-sampling methods, including RS, SMOTE, and OCVAE, which means that the samples augmented by MGVAE have much less noise. Besides, we find that the MGVAE achieves a larger lead under the imbalance ratio (IR) $\rho = 600$. For example, the percentage improvement of the GM term compared to SMOTE is $\Delta = 5.8$ and $\Delta = 13.2$, respectively. This means that our method remains effective in the case of extreme imbalance.

**CelebA.** Then, we move on to CelebA, a much larger multi-label celebrity face dataset with a resized resolution of 64*64*3. We downsample the dataset based on hair color to obtain a 5-class long-tailed dataset with an imbalance rate of $\rho = N1/N5 = 100$. For classifier selection, we train a ResNet-20 (He et al., 2016) using cross-entropy loss for 90 epochs with mini-batch size 100. The learning rate is initialized to 0.01 with a weight decay of $2e$-4 and an exponential decrease of $\gamma = 0.95$. In addition, all the generative models, including MGVAE, OCVAE, and OCDCGAN, introduce convolution layers, which are more suitable for complex image generation. Meanwhile, the main architecture of the models is kept consistent. The EWC regularization weight $\lambda$ in MGVAE is set to 50. To pursue uniformity, we only compare MGVAE with the over-sampling-based methods in this section.

Table 4: Comparison of classification performance on the long-tail CelebA.

| Methods | B-ACC | ASCA | GM |
|---|---|---|---|
| ERM | $62.3 \pm 0.9$ | $63.8 \pm 0.5$ | $40.2 \pm 0.3$ |
| RS | $64.2 \pm 0.5$ | $65.6 \pm 0.4$ | $45.0 \pm 1.2$ |
| SMOTE | $63.5 \pm 0.8$ | $65.0 \pm 0.6$ | $42.2 \pm 0.8$ |
| OCVAE | $64.4 \pm 0.8$ | $65.5 \pm 1.0$ | $48.5 \pm 0.6$ |
| OCDCGAN | $65.8 \pm 0.1$ | $67.2 \pm 0.0$ | $52.4 \pm 1.0$ |
| **MGVAE (ours)** | **66.8** $\pm 0.2$ | **68.0** $\pm 0.2$ | **55.6** $\pm 0.2$ |

The results are shown in Table 4. Once again, our method outperforms all other baseline methods, demonstrating that the minority samples augmented by MGVAE are more diverse and meaningful. Remarkably, as a VAE-based method, MGVAE surpasses OCDCGAN, further showing our algorithm's effectiveness.

**Tabular Data.** Finally, we test our model on several real-world tabular datasets, namely Musk, Water Quality, and Isolet. All these datasets are naturally imbalanced, as detailed in Table 1. For compatibility with the generative model, we preprocess all the tabular data to the scale of [-1,1] or [0,1] by dimension according to the sign of the original data. We train a 2-layer fully-connected neural network for 100 epochs with mini-batch size 100, and the learning rate is $1e$-3. The EWC regularization weight of MGVAE is 500.

The results of the three datasets are summarized in Tables 5, 6, and 7, respectively. Our method outperforms the other

Table 5: Comparison of classification performance on the Musk.

| Methods | B-ACC | ASCA | GM |
|---|---|---|---|
| ERM | 50.0 ± 0.0 | 50.0 ± 0.0 | 0.0 ± 0.0 |
| FOCAL | 87.1 ± 1.5 | 87.2 ± 1.6 | 86.1 ± 1.4 |
| RW | 85.6 ± 1.1 | 86.7 ± 1.0 | 84.3 ± 1.2 |
| CBRW | 84.2 ± 0.5 | 84.3 ± 0.5 | 82.7 ± 0.6 |
| LDAM | 89.7 ± 1.2 | 89.8 ± 1.2 | 89.0 ± 1.3 |
| RS | 84.9 ± 0.9 | 85.0 ± 1.0 | 83.6 ± 1.1 |
| SMOTE | 83.6 ± 1.3 | 83.5 ± 1.4 | 81.8 ± 1.7 |
| OCVAE | 90.5 ± 0.3 | 90.1 ± 0.2 | 89.8 ± 0.2 |
| **MGVAE (ours)** | **92.4** ± 1.2 | **92.4** ± 1.3 | **92.1** ± 1.4 |

Table 6: Comparison of classification performance on the Water Quality.

| Methods | B-ACC | ASCA | GM |
|---|---|---|---|
| ERM | 50.0 ± 0.0 | 50.0 ± 0.0 | 0.0 ± 0.0 |
| FOCAL | 65.7 ± 0.7 | 65.7 ± 0.7 | 63.3 ± 1.1 |
| RW | 57.7 ± 0.1 | 57.6 ± 0.3 | 48.7 ± 0.2 |
| CBRW | 56.4 ± 0.2 | 56.4 ± 0.3 | 49.5 ± 0.6 |
| LDAM | 65.8 ± 0.3 | 65.9 ± 0.2 | 62.0 ± 0.3 |
| RS | 58.3 ± 1.5 | 58.2 ± 1.4 | 51.3 ± 1.1 |
| SMOTE | 56.8 ± 1.4 | 56.0 ± 1.5 | 51.9 ± 1.6 |
| OCVAE | 61.9 ± 1.2 | 62.1 ± 1.9 | 58.3 ± 1.5 |
| **MGVAE (ours)** | **67.2** ± 1.1 | **67.3** ± 1.3 | **65.9** ± 1.2 |

Table 7: Comparison of classification performance on the Isolet.

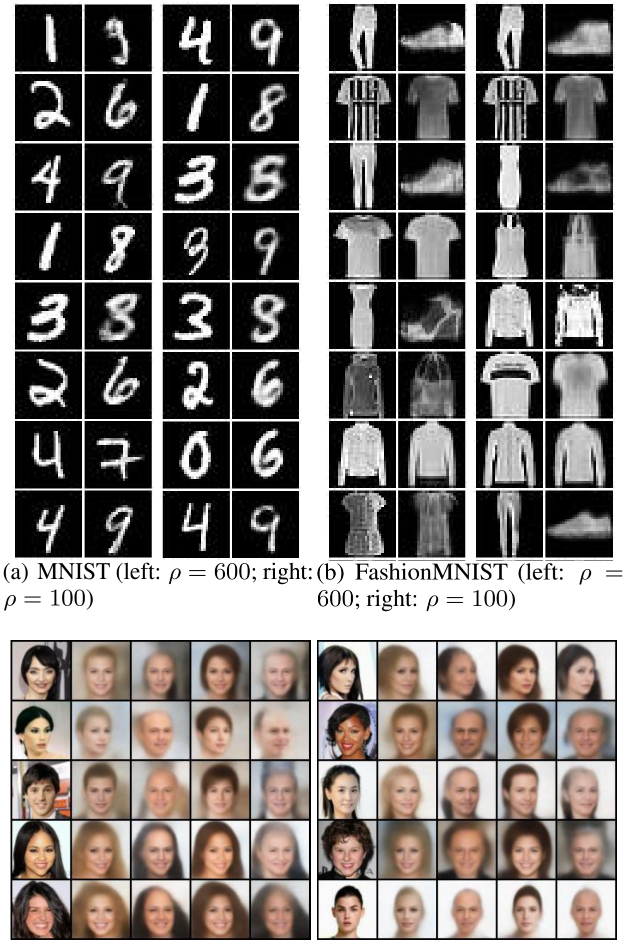
| Methods | B-ACC | ASCA | GM |
|---|---|---|---|
| ERM | 50.0 ± 0.0 | 50.0 ± 0.0 | 0.0 ± 0.0 |
| FOCAL | 50.0 ± 0.0 | 50.0 ± 0.0 | 0.0 ± 0.0 |
| RW | 92.9 ± 0.4 | 92.9 ± 0.2 | 92.7 ± 0.2 |
| CBRW | 92.8 ± 0.4 | 92.8 ± 0.5 | 92.6 ± 0.6 |
| LDAM | 93.5 ± 1.0 | 93.5 ± 1.0 | 93.3 ± 1.1 |
| RS | 92.9 ± 0.3 | 92.8 ± 0.4 | 92.7 ± 0.4 |
| SMOTE | 93.8 ± 0.7 | 93.7 ± 0.8 | 93.6 ± 0.6 |
| OCVAE | 91.6 ± 0.4 | 91.7 ± 0.2 | 91.5 ± 0.2 |
| **MGVAE (ours)** | **95.1** ± 0.6 | **95.0** ± 0.7 | **94.9** ± 0.7 |

methods in all metrics, proving the effectiveness of our model on real-world data.

### 3.1.2 Qualitative Results

To further demonstrate the effectiveness of our proposed model, we also visualize the generated minority samples that are desirably supposed to achieve the diversity and richness of the majority ones while maintaining self-semantic information.

Firstly, we visualize the generated minority samples of our model in different datasets according to the generation process of MGVAE (see Algorithm 1). The results are shown in Figure 3. In overall, MGVAE can generate meaningful and distinguishable samples for different datasets and imbalance ratios, i.e., no model collapse. Since the generation process of MGVAE is in a one-to-one mapping style from majority to minority, we show both the reference majority sample and the corresponding minority one. In each plate of MNIST and FashionMNIST in Figure 3, the left column is the reference majority sample, and the right columns are the generated minority ones correspondingly. For CelebA, the first column is the reference majority sample from the black hair. The other columns are the corresponding generated samples of the minority classes, i.e., from left to right, blond hair, bald, brown hair, and gray hair.



(a) MNIST (left: $\rho = 600$; right: $\rho = 100$)

(b) FashionMNIST (left: $\rho = 600$; right: $\rho = 100$)

(c) CelebA (from left to right: black,blonde,bald,brown,and gray)

Figure 3: Samples from MGVAE. In each plate, the first column is the reference majority sample, and the rest columns are the corresponding generated minority ones. Best viewed in color.

Form Figure 3, we notice that the generated minority samples have great stylistic similarity with their reference majority samples, while the semantic information is different. For example, the number writing style in MNIST, such as tilt angle and thickness, remains the same. Some features other

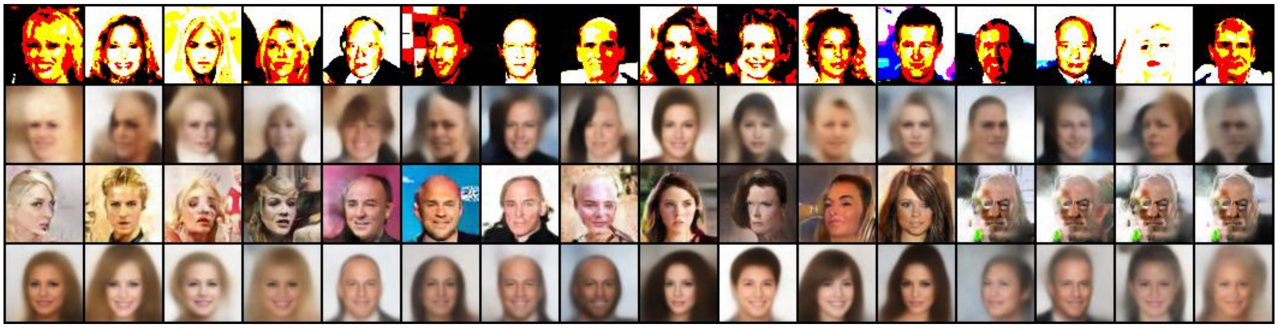


Figure 4: Comparison of the generation of different methods. Each row corresponds to one method, from top to bottom: SMOTE, OCVAE, OCDCGAN, and MGVAE. Each group of the four columns corresponds to one minority class, from left to right: blonde hair, bald hair, brown hair, and gray hair.

than hair color in CelebA, such as the face orientation and mouth shape, are consistent. In other words, the minority augmentation samples inherit the diversity of the majority ones in the process of generation. Moreover, we can see that the quality of the generated images improves as the scale of the minority increases, i.e., the imbalance ratio decreases.

Secondly, We compare the sampling quality of all oversampling-based methods, and the results on CelebA are shown in Figure 4. More results on MNIST and FashionMNIST are presented in the supplementary materials D.1 because of the space constraints. As shown in Figure 4, we find that the samples generated by SMOTE differ greatly from the original dataset, and the sampled images of OCVAE are very blurry and have no obvious semantic features. OCDCGAN could create sharper but severely distorted images, and the model collapses when the sample size is small (See the last four columns in the OCDCGAN row). Compared with the above methods, MGVAE can maintain the semantic information of each class while generating clear images without model collapse.

### 3.2 Ablation study

To gain a deeper understanding of the role played by each part of our model, we next perform a set of ablation experiments. Specifically, we first replace the prior with Gaussian to verify the effectiveness of the majority-based prior of MGVAE, i.e., we test on the standard VAE with the pretrain-finetuning training process and the EWC regularization. Second, we incrementally remove the EWC regularization and the pretrain-finetuning training process in MGVAE. The corresponding results in MNIST are summarised in Table 8, from which we know that these two parts can significantly improve the model's effectiveness, especially the pretrain-finetuning process. The reason is that they can avoid overfitting and model collapse by better incorporating and preserving the information of the majority class.

Table 8: Ablation study of the classification performance in MNIST.

| IR | $\rho = 100$ | | | $\rho = 600$ | | |
|---|---|---|---|---|---|---|
| Methods | B-ACC | ASCA | GM | B-ACC | ASCA | GM |
| VAE w/ PT+EWC | $79.1 \pm 0.4$ | $78.4 \pm 0.6$ | $75.7 \pm 0.4$ | $60.3 \pm 0.8$ | $59.3 \pm 0.7$ | $42.7 \pm 0.3$ |
| MGVAE w/o PT | $81.1 \pm 0.6$ | $80.6 \pm 0.5$ | $78.1 \pm 0.7$ | $59.8 \pm 0.8$ | $58.8 \pm 0.5$ | $41.1 \pm 1.0$ |
| MGVAE w/o EWC | $84.0 \pm 0.7$ | $83.5 \pm 0.5$ | $81.9 \pm 0.9$ | $62.5 \pm 0.7$ | $61.5 \pm 0.4$ | $47.7 \pm 1.2$ |
| **MGVAE** | **85.0** $\pm 0.2$ | **84.6** $\pm 0.2$ | **83.2** $\pm 0.2$ | **65.4** $\pm 1.0$ | **64.4** $\pm 1.1$ | **53.4** $\pm 1.1$ |

### 3.3 Sensitivity Analysis on $\lambda$

In our model, we employ Elastic Weight Consolidation (EWC) regularization during the fine-tuning process to prevent catastrophic forgetting of the majority and overfitting to the minority. The corresponding hyper-parameter is $\lambda$. To analyze the impact of $\lambda$, we conduct experiments on FashionMNIST with two different imbalanced ratios. The range of values of $\lambda$ is $\{5e2, 5e4, 5e6, 5e8\}$. And the results are summarized in Table 9, and each averaged from three random trials.

Table 9: Comparison of classification performance of MGVAE with different $\lambda$ on FashionMNIST.

| IR | $\rho = 100$ | | | $\rho = 600$ | | |
|---|---|---|---|---|---|---|
| $\lambda$ | B-ACC | ASCA | GM | B-ACC | ASCA | GM |
| 5e2 | $88.3 \pm 0.1$ | $88.3 \pm 0.1$ | $87.5 \pm 0.1$ | $84.1 \pm 0.6$ | $84.1 \pm 0.5$ | $82.8 \pm 0.4$ |
| 5e4 | $88.3 \pm 0.1$ | $88.4 \pm 0.1$ | $87.6 \pm 0.1$ | $84.8 \pm 0.4$ | $84.8 \pm 0.4$ | $83.6 \pm 0.4$ |
| 5e6 | $88.0 \pm 0.3$ | $88.1 \pm 0.2$ | $87.2 \pm 0.3$ | $84.0 \pm 0.5$ | $84.0 \pm 0.5$ | $82.4 \pm 0.7$ |
| 5e8 | $87.7 \pm 0.3$ | $87.7 \pm 0.3$ | $87.0 \pm 0.5$ | $83.8 \pm 0.6$ | $83.8 \pm 0.6$ | $82.0 \pm 0.9$ |

## 4 Related Work

Over-sampling has been widely used to solve class-imbalanced issues with the emergence of many advanced methods in recent decades. Generally, over-sampling methods provide a way to augment the minority class information, resulting in a class-balanced dataset for the downstream tasks. Random over-sampling (ROS) (Japkowicz, 2000) is one of the straightforward ways to re-balance class by repeatedly sampling the minority samples. However, ROS does not inherently have any information augmentation and is prone to overfitting in the minority classes. SMOTE (Chawla et al., 2002) augments the minority class with new instances generated by interpolating neighboring minority class instances to address this issue. After that, some following work (Han et al., 2005; He et al., 2008; Mullick et al., 2019) based on SMOTE was proposed to improve the performance. However, the synthesized samples are usually noisy due to the boundary samples, especially for image data. Besides, the performance drops drastically under the extreme imbalance ratio. Regardless of ROS or SMOTE-based methods, only instances of the minority class are used in data augmentation, which is small-scale and easy to overfit. Unlike these, major-to-minor translation (M2m) (Kim et al., 2020) augments minority classes by translating the majority ones according to the trained classifier under imbalanced data, which is essentially the generation process of adversarial examples. Therefore, the augmented minority sample generated by M2m is a slight perturbation of the majority one, which is visually contrary to human cognition. More recently, CMO (Park et al., 2022) augments the minority by leveraging the rich context of the majority classes as background images, achieved by CutMix, while the generated minority sample is semantically meaningless. Another research line of over-sampling is related to deep generative models (DGMs), such as Conditional VAE (Fajardo et al., 2021), Contrastive VAE (Dai et al., 2019), GAN (Pourreza et al., 2021; Mullick et al., 2019), etc. These models can generate corresponding minority samples based on the class label. But, the model will collapse on minority class when the sample scale tends to a few. Our proposed model, MGVAE, is also a generative model, inheriting the architecture of VAE. Specifically, MGVAE generates the minority sample according to a majority-based prior, resulting in one-to-one sample mapping. The generated samples are meaningful, and class information is consistent with human cognition.

## 5 Conclusion

In the paper, we introduce Majority-Guided VAE (MGVAE), a novel over-sampling model to re-balance datasets by generating the new minority samples based on the majority prior. By utilizing the information of the majority class, MGVAE can generate the minority samples with better diversity and richness. As a result, overfitting can be relatively avoided in downstream tasks based on the augmented data. Additionally, to better control the training process of the model, we use pretrain-finetune two-stage training and Elastic Weight Consolidation (EWC) regularization. Finally, we demonstrate the promising performance of the MGVAE in the downstream classification task on both benchmark image datasets and real-world tabular data.

## Limitation and Societal impact

At present, MGVAE still has some limitations. First, MGVAE is essentially a variant of VAE. Although the generation quality is much improved compared to the standard VAE, the generated images are still blurred compared to other advanced generation models such as Normalizing Flows, Diffusion models, etc. Second, in the current form of MGVAE, only the majority data is used during the pre-train process, resulting in a majority-based prior for our minority generative. We would like to test the different settings for pre-train data selection in our future work. For potential societal impact, MGVAE may be used to generate fake pictures that are not present in reality, such as human faces.

### Acknowledgements

This work was partially supported by the National Key Research and Development Program of China (No. 2018AAA0100204), a key program of fundamental research from Shenzhen Science and Technology Innovation Commission (No. JCYJ20200109113403826), the Major Key Project of PCL (No. PCL2021A06), an Open Research Project of Zhejiang Lab (NO.2022RC0AB04), and Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies (No. 2022B1212010005).

### References

Ai, Q., He, L., Liu, S., and Xu, Z. (2021). Bype-vae: Bayesian pseudocoresets exemplar vae. *Advances in Neural Information Processing Systems*, 34:5910–5920.

Asuncion, A. and Newman, D. (2007). Uci machine learning repository.

Branco, P., Torgo, L., and Ribeiro, R. P. (2016). A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys (CSUR)*, 49(2):1–50.

Cao, K., Wei, C., Gaidon, A., Arechiga, N., and Ma, T. (2019). Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32.

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

Chung, Y.-A., Lin, H.-T., and Yang, S.-W. (2015). Cost-aware pre-training for multiclass cost-sensitive deep learning. *arXiv preprint arXiv:1511.09337*.

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*, pages 9268–9277.

Cui, Y., Song, Y., Sun, C., Howard, A., and Belongie, S. (2018). Large scale fine-grained categorization and domain-specific transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4109–4118.

Dai, W., Ng, K., Severson, K., Huang, W., Anderson, F., and Stultz, C. (2019). Generative oversampling with a contrastive variational autoencoder. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 101–109. IEEE.

Das, S., Datta, S., and Chaudhuri, B. B. (2018). Handling data irregularities in classification: Foundations, trends, and future challenges. *Pattern Recognition*, 81:674–693.

Dinh, L., Krueger, D., and Bengio, Y. (2014). Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*.

Dong, Q., Gong, S., and Zhu, X. (2018). Imbalanced deep learning by minority class incremental rectification. *IEEE transactions on pattern analysis and machine intelligence*, 41(6):1367–1381.

Fajardo, V. A., Findlay, D., Jaiswal, C., Yin, X., Houmanfar, R., Xie, H., Liang, J., She, X., and Emerson, D. (2021). On oversampling imbalanced data with deep conditional generative models. *Expert Systems with Applications*, 169:114463.

Fan, G., Zhang, C., Chen, J., Li, B., Xu, Z., Li, Y., Peng, L., and Gong, Z. (2022). Field-aware variational autoencoders for billion-scale user representation learning. In *38th IEEE International Conference on Data Engineering, ICDE 2022, Kuala Lumpur, Malaysia, May 9-12, 2022*, pages 3413–3425. IEEE.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Han, H., Wang, W.-Y., and Mao, B.-H. (2005). Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pages 878–887. Springer.

He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328. IEEE.

He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851.

Huang, C., Li, Y., Loy, C. C., and Tang, X. (2016). Learning deep representation for imbalanced classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5375–5384.

Huang, C., Li, Y., Loy, C. C., and Tang, X. (2019). Deep imbalanced learning for face recognition and attribute prediction. *IEEE transactions on pattern analysis and machine intelligence*, 42(11):2781–2794.

Japkowicz, N. (2000). The class imbalance problem: Significance and strategies. In *Proc. of the Int'l Conf. on Artificial Intelligence*, volume 56, pages 111–117. Citeseer.

Johnson, J. M. and Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54.

Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., and Kalantidis, Y. (2019). Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*.

Khan, S. H., Hayat, M., Bennamoun, M., Sohel, F. A., and Togneri, R. (2017). Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE transactions on neural networks and learning systems*, 29(8):3573–3587.

Kim, J., Jeong, J., and Shin, J. (2020). M2m: Imbalanced classification via major-to-minor translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13896–13905.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Kubat, M., Matwin, S., et al. (1997). Addressing the curse of imbalanced training sets: one-sided selection. In *Icml*, volume 97, page 179. Citeseer.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Li, Y., Zhang, R., Lu, J. C., and Shechtman, E. (2020). Few-shot image generation with elastic weight consolidation. In *Advances in Neural Information Processing Systems*.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Liu, X.-Y., Wu, J., and Zhou, Z.-H. (2008). Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550.

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

Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., and Van Der Maaten, L. (2018). Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 181–196.

Mullick, S. S., Datta, S., and Das, S. (2019). Generative adversarial minority oversampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1695–1704.

Nguyen, C. V., Li, Y., Bui, T. D., and Turner, R. E. (2017). Variational continual learning. *arXiv preprint arXiv:1710.10628*.

Ojha, U., Li, Y., Lu, J., Efros, A. A., Lee, Y. J., Shechtman, E., and Zhang, R. (2021). Few-shot image generation via cross-domain correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10743–10752.

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*, pages 6887–6896.

Pourreza, M., Mohammadi, B., Khaki, M., Bouindour, S., Snoussi, H., and Sabokrou, M. (2021). G2d: generate to detect anomaly. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2003–2012.

Ren, M., Zeng, W., Yang, B., and Urtasun, R. (2018). Learning to reweight examples for robust deep learning. In *International conference on machine learning*, pages 4334–4343. PMLR.

Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., and Belongie, S. (2018). The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778.

Wang, C., Gao, S., Wang, P., Gao, C., Pei, W., Pan, L., and Xu, Z. (2022). Label-aware distribution calibration for long-tailed classification. *IEEE Transactions on Neural Networks and Learning Systems*.

Wang, T., Li, Y., Kang, B., Li, J., Liew, J., Tang, S., Hoi, S., and Feng, J. (2020a). The devil is in classification: A simple framework for long-tail instance segmentation. In *European conference on computer vision*, pages 728–744. Springer.

Wang, X., Lyu, Y., and Jing, L. (2020b). Deep generative model for robust imbalance classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14124–14133.

Wang, Y., Wu, C., Herranz, L., van de Weijer, J., Gonzalez-Garcia, A., and Raducanu, B. (2018). Transferring gans: generating images from limited data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 218–234.

Wang, Y.-X., Ramanan, D., and Hebert, M. (2017). Learning to model the tail. *Advances in Neural Information Processing Systems*, 30.

Xiao, H., Rasul, K., and Vollgraf, R. (2017). Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

Zhang, Y., Kang, B., Hooi, B., Yan, S., and Feng, J. (2021). Deep long-tailed learning: A survey. *arXiv preprint arXiv:2110.04596*.

Zhang, Z. and Pfister, T. (2021). Learning fast sample re-weighting without reward data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 725–734.

# A Derivation of Eq.(4)

Following the definition in Section Methods 2, we have

$$
\begin{aligned}
&\log p(\mathbf{x}^- \mid \mathcal{X}^+, \Psi) \\
&= \log \sum_{n=1}^{N^+} \frac{1}{N^+} T_\Psi(\mathbf{x}^- \mid \mathbf{x}_n^+) \\
&= \log \sum_{n=1}^{N^+} \frac{1}{N^+} \int_z r_\phi(\mathbf{z} \mid \mathbf{x}_n^+) p_\theta(\mathbf{x}^- \mid \mathbf{z}) \mathbf{dz} \\
&= \log \int_z p_\theta(\mathbf{x}^- \mid \mathbf{z}) \sum_{n=1}^{N^+} \frac{1}{N^+} r_\phi(\mathbf{z} \mid \mathbf{x}_n^+) \mathbf{dz} \\
&= \log \int_z \frac{q_\phi(\mathbf{z} \mid \mathbf{x}^-) p_\theta(\mathbf{x}^- \mid \mathbf{z}) \sum_{n=1}^{N^+} r_\phi(\mathbf{z} \mid \mathbf{x}_n^+)/N^+}{q_\phi(\mathbf{z} \mid \mathbf{x}^-)} \mathbf{dz} \\
&= \log \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}^-)} \left[ p_\theta(\mathbf{x}^- \mid \mathbf{z}) \frac{\sum_{n=1}^{N^+} r_\phi(\mathbf{z} \mid \mathbf{x}_n^+)/N^+}{q_\phi(\mathbf{z} \mid \mathbf{x}^-)} \right] \\
&\geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}^-)} \log \left[ p_\theta(\mathbf{x}^- \mid \mathbf{z}) \frac{\sum_{n=1}^{N^+} r_\phi(\mathbf{z} \mid \mathbf{x}_n^+)/N^+}{q_\phi(\mathbf{z} \mid \mathbf{x}^-)} \right] \\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}^-)} \log p_\theta(\mathbf{x}^- \mid \mathbf{z}) - \log \frac{q_\phi(\mathbf{z} \mid \mathbf{x}^-)}{\sum_{n=1}^{N^+} r_\phi(\mathbf{z} \mid \mathbf{x}_n^+)/N^+} \\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}^-)} \log p_\theta(\mathbf{x}^- \mid \mathbf{z}) - \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}^-)} \log \frac{q_\phi(\mathbf{z} \mid \mathbf{x}^-)}{\sum_{n=1}^{N^+} r_\phi(\mathbf{z} \mid \mathbf{x}_n^+)/N^+} \\
&\equiv O(\Psi, \mathcal{X}^+; \mathbf{x}^-),
\end{aligned}
$$

where the inequality relation is derived from the Jensen's inequality.

# B Loss re-weighting strategies

We compare MGVAE against other different loss re-weighting strategies in the experimental section, namely FOCAL and LDAM loss. For the sake of completeness of the content, we briefly introduce these loss re-weighting strategies here.

## B.1 FOCAL Loss

To address the imbalance problem in object detection, the Focal loss is introduced to balance the sample-wise classification loss for model training by down-weighing easy samples. In detail, the Focal loss adds a re-weighting factor $(1 - h_i)^\gamma$ with $\gamma > 0$ to the standard cross-entropy loss $\mathcal{L}_{CE}$, where $h_i$ is a probability prediction for the sample $x_i$ over its true category $y_i$. The resultant loss takes the form as follows:

$$\mathcal{L}_{Focal} = (1 - h_i)^\gamma \mathcal{L}_{CE} = -(1 - h_i)^\gamma \log(h_i). \tag{7}$$

As a result, the cross-entropy loss of the easy samples, which may dominate the training by the large predicted probability $h_i$ for their true categories, will be down-weighted. Note that the $\gamma$ is set to 1.0 in all our experiments.

## B.2 LDAM Loss

The label-distribution-aware margin (LDAM) loss expands the decision boundaries of few-shot classes, resulting in larger margins between those classes. The final loss is formulated as a cross-entropy loss with enforced margins:

$$\mathcal{L}_{\text{LDAM}} := -\log \frac{e^{\hat{y}_j - \Delta_j}}{e^{\hat{y}_j - \Delta_j} + \sum_c \neq j e^{\hat{y}_c - \Delta_c}}, \tag{8}$$

where $\hat{y}$ are the logits and $\Delta_j$ is a class-aware margin, inversely proportional to $n_j^{1/4}$, and $n$ is the number of class samples.

# C Detailed Training Process

To better understand our method, we summarize the detailed training process in Algorithm 2.

**Algorithm 2:** The Training Process of MGVAE.

```
Input  : Training class-imbalanced data $\mathcal{D}_{imb} = \{\mathcal{X}^+ \equiv \{\mathbf{x}_n^+\}_{n=1}^{N^+}, \mathcal{X}^- \equiv \{\mathbf{x}_n^-\}_{n=1}^{N^-}$, where $N^+ \gg N^-$.
         Batch size $B$.
         Pre-train steps $T_1$ and fine-tune steps $T_2$.
         Majority down-sample size $S$.
         Decoder $p_\theta$, encoder $q_\phi$, and the prior $r_\phi$.
Output : Parameters $\theta$ and $\phi$ of generative model $M$ for minority.
/* Pre-training                                                    */
for $t = 1, \cdots, T_1$ do
    Obtain a mini-batch of $B$ majority datapoints
    Randomly down-sample $S$ majority samples $\{\mathbf{x}_n\}_{n=1}^{S}$for computing prior $r_\phi$
     $\mathcal{L}_{\text{pretrain}} = \frac{1}{B}\sum_{i=1}^{B}\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_i)}\left[\log p_\theta(\mathbf{x}_i \mid \mathbf{z}) - \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_i)}\log\frac{q_\phi(\mathbf{z}|\mathbf{x}_i)}{\sum_{n=1}^{S} r_\phi(\mathbf{z}|\mathbf{x}_n)/S}\right]$
    Optimize $\mathcal{L}_{\text{pretrain}}$ with ADAM optimizer.
Get pre-trained model $\mathcal{M}_{\text{pretrain}}$.
/* Fine-tuning                                                     */
for $t = 1, \cdots, T_2$ do
    Obtain a mini-batch of $B$ minority datapoints
    Randomly down-sample $S$ majority samples $\{\mathbf{x}_n\}_{n=1}^{S}$for computing prior $r_\phi$
     Get Fisher Information vector $F$ of model $M_{pretrained}$ by Eq. 5
     Optimize optimization objective $O_{EWC}$ in Eq. 6 with ADAM optimizer.
Get trained model $\mathcal{M}$.
```

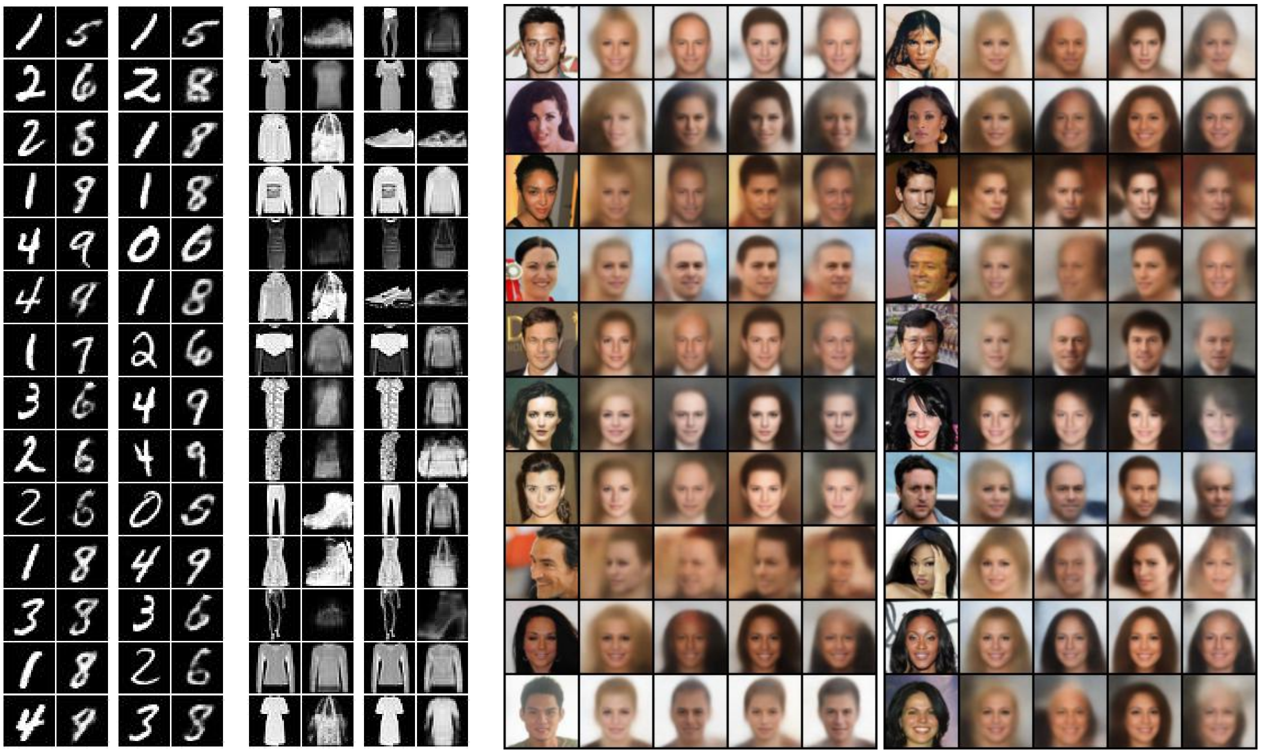
# D More experimental results

## D.1 More visual results

Due to space constraints, we give part of the generative samples in the Qualitative Results of the experimental section. To demonstrate the effectiveness of MGVAE, we provide more visual results.

**Samples from MGVAEs** First, more sampling results on different models and datasets are presented in Figure 5. Consistent with the main body, the left column of each plate of MNIST and FashionMNIST is the exemplar majority sample, and the right columns are the generated minority ones correspondingly. For Celeba, the first column is the exemplar majority sample from the black hair. The other columns are the corresponding generated samples of the minority classes, from left to right, blond hair, bald, brown hair, and gray hair.

**Comparison results** In the main body of the paper, we only give the comparison sampling results of different methods on Celeba due to space constraints. More results are shown in Figure 6. Compared with the other methods, MGVAE can maintain the semantic information of each class while generating clear images without model collapse.

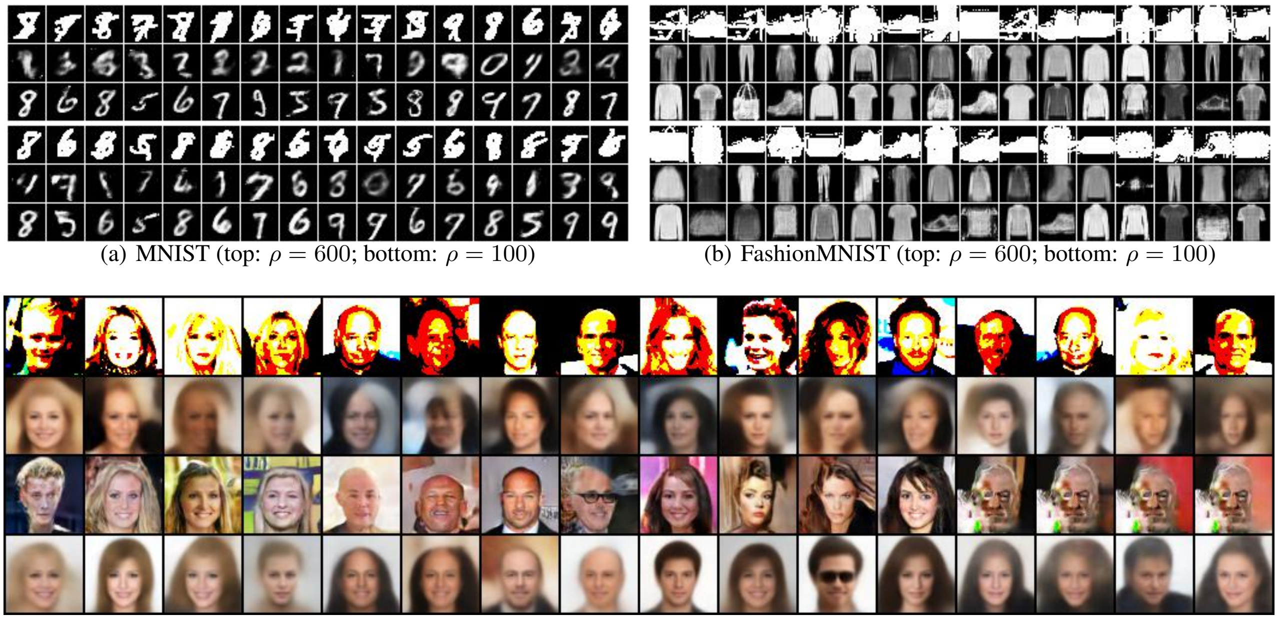


(a) MNIST (left: $\rho$ = 600; right: $\rho = 100$)

(b) FashionMNIST (left: $\rho$ = 600; right: $\rho$ = 100)

(c) CelebA (from left to right: black,blonde,bald,brown,and gray)

Figure 5: Samples from MGVAE.



(a) MNIST (top: $\rho = 600$; bottom: $\rho = 100$)

(b) FashionMNIST (top: $\rho = 600$; bottom: $\rho = 100$)

(c) CelebA (from left to right: black,blonde,bald,brown,and gray)

Figure 6: Comparison of the generation of different methods. Each row corresponds to one method. For MNIST and FashionMNIST, from top to bottom: SMOTE, OCVAE, and MGVAE. For Celeba, from top to bottom: SMOTE, OCVAE, OCDCGAN, and MGVAE. Each group of the four columns corresponds to one minority class, from left to right: blonde hair, bald hair, brown hair, and gray hair.

### D.2 Upper bounds

For experimental completeness, we provide the results of the used classifiers on the fully-balanced MNIST and FashionMNIST, which can be considered the classification upper bounds. The results are shown in Table 10.

Table 10: Classification results on fully-balanced MNIST and FashionMNIST.

| MNIST | | | FashionMNIST | | |
|---|---|---|---|---|---|
| B-ACC | ASCA | GM | B-ACC | ASCA | GM |
| $98.5 \pm 0.0$ | $98.5 \pm 0.0$ | $98.5 \pm 0.0$ | $94.3 \pm 0.1$ | $94.3 \pm 0.1$ | $94.2 \pm 0.1$ |

## E Architectures

### E.1 Architectures of the DGMs

For a fair comparison, the structure of all generated models is kept consistent for each particular dataset. Specifically, the neural network uses fully-connected layers (denoted MLP) for the relatively simple datasets MNIST, FashionMNIST, and Tabular data. For Celeba, a convolutional layer (denoted CNN) is used. Note that the architecture of the convolutional layer is based on this code repository[1]. We use curly brackets to denote concatenation; the number in a bracket means the layer size.

**VAE-based architecture with MLP:** including MGVAE and OCVAE for MNIST, FashionMNIST, and Tabular data.

$$\begin{aligned}
&\textbf{Encoder:} \\
&\qquad \text{E1} = \text{MLP [Input dim} - \text{Hidden dim 1]} \\
&\qquad \text{E2} = \text{MLP [Hidden dim 1} - \text{Hidden dim 2]} \\
&\qquad \log \sigma^2 = \text{MLP [Hidden dim 2} - \text{Latent dim]} \\
&\qquad \mu_z = \text{MLP [Hidden dim 2} - \text{Latent dim]} \\
&\textbf{Decoder:} \\
&\qquad \text{D1} = \text{MLP [Latent dim} - \text{Hidden dim 2]} \\
&\qquad \text{D2} = \text{MLP [Hidden dim 2} - \text{Hidden dim 1]} \\
&\qquad \mu_x = \text{MLP [Hidden dim 2} - \text{Output dim]}
\end{aligned}$$

Specifically, for MNIST and FashionMNIST, Input dim=Output dim=784, Hidden dim 1=Hidden dim 2=300, and Latent dim=40. For Tabular data, Input dim=Output dim=Data dim, Hidden dim 1=Hidden dim 2=300, and Latent dim=10.

[1] https://github.com/sajadn/Exemplar-VAE/blob/master/models/fully_conv.py

**VAE-based architecture with CNN:** including MGVAE and OCVAE for Celeba.

**Encoder:**

$$
\begin{aligned}
\mathrm{E1} &= \mathrm{CNN}\ [64 \times 64 \times 3 - 32 \times 32 \times 64] \\
\mathrm{E2} &= \mathrm{CNN}\ [32 \times 32 \times 64 - 16 \times 16 \times 128] \\
\mathrm{E3} &= \mathrm{CNN}\ [16 \times 16 \times 128 - 8 \times 8 \times 256] \\
\mathrm{E4} &= \mathrm{CNN}\ [8 \times 8 \times 256 - 4 \times 4 \times 512] \\
\log \sigma^2 &= \mathrm{MLP}\ [4 \times 4 \times 512 - \text{Latent dim}] \\
\mu_z &= \mathrm{MLP}\ [4 \times 4 \times 512 - \text{Latent dim}]
\end{aligned}
$$

**Decoder:**

$$
\begin{aligned}
\mathrm{D1} &= \mathrm{MLP}\ [\text{Latent dim} - 4 \times 4 \times 512] \\
&\quad \text{Upsample(2)} \\
\mathrm{D2} &= \mathrm{CNN}\ [8 \times 8 \times 512 - 16 \times 16 \times 256] \\
\mathrm{D3} &= \mathrm{CNN}\ [16 \times 16 \times 256 - 32 \times 32 \times 128] \\
\mathrm{D4} &= \mathrm{CNN}\ [32 \times 32 \times 128 - 64 \times 64 \times 64] \\
\mu_x &= \mathrm{CNN}\ [64 \times 64 \times 64 - 64 \times 64 \times 3]
\end{aligned}
$$

We use Latent dim=40 for all our models in Celeba, including MGVAE and OCVAE.

**Architecture of OCDCGAN:** including OCDCGAN for Celeba.

**Generator:**

$$
\begin{aligned}
\mathrm{G1} &= \mathrm{CNN}\ [\text{Latent dim} - 4 \times 4 \times 1024] \\
\mathrm{G2} &= \mathrm{CNN}\ [4 \times 4 \times 1024 - 8 \times 8 \times 512] \\
\mathrm{G3} &= \mathrm{CNN}\ [8 \times 8 \times 512 - 16 \times 16 \times 256] \\
\mathrm{G4} &= \mathrm{CNN}\ [16 \times 16 \times 256 - 32 \times 32 \times 128] \\
\mathrm{G5} &= \mathrm{CNN}\ [32 \times 32 \times 128 - 64 \times 64 \times 3]
\end{aligned}
$$

**Discriminator:**

$$
\begin{aligned}
\mathrm{G1} &= \mathrm{CNN}\ [64 \times 64 \times 3 - 32 \times 32 \times 128] \\
\mathrm{G2} &= \mathrm{CNN}\ [32 \times 32 \times 128 - 16 \times 16 \times 256] \\
\mathrm{G3} &= \mathrm{CNN}\ [16 \times 16 \times 256 - 8 \times 8 \times 512] \\
\mathrm{G4} &= \mathrm{CNN}\ [8 \times 8 \times 512 - 4 \times 4 \times 1024] \\
\mathrm{G5} &= \mathrm{CNN}\ [4 \times 4 \times 1024 - 1 \times 1 \times 1] \\
&\quad \text{Sigmoid()}
\end{aligned}
$$

The Latent dim is 100 in our setting.

### E.2 Architectures of the Classifiers

Similarly, for the classifier, we also use two network structures: fully-connected layers for MNIST, FashionMNIST, Tabular data, and ResNet-20 for Celeba. The architecture of the fully-connected network is as follows. See our code implementation for details.

**MLP classifier.**

$$
\begin{aligned}
\mathrm{C1} &= \mathrm{MLP}\ [\text{Input dim} - 256] \\
\mathrm{C2} &= \mathrm{MLP}\ [256 - 128] \\
\mathrm{C3} &= \mathrm{MLP}\ [128 - \text{Class num}]
\end{aligned}
$$

The architecture of the ResNet-20 network is based on this implementation[2].

[2] https://github.com/akamaster/pytorch_resnet_cifar10/blob/master/resnet.py