MINING MULTI-LABEL SAMPLES FROM SINGLE POSITIVE LABELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Conditional generative adversarial networks (cGANs) have shown superior results in class-conditional generation tasks. In order to simultaneously control multiple conditions, cGANs requires multi-label training datasets, where multiple labels can be assigned to each data instance. Nevertheless, the tremendous annotation cost limits the accessibility of multi-label datasets in the real world. Hence, we explore the practical setting called *single positive setting*, where each data instance is annotated by only one positive label with no explicit negative labels. To generate multi-label data in the single positive setting, we propose a novel sampling approach called single-to-multi-label (S2M) sampling, based on the Markov chain Monte Carlo method. As a widely applicable "add-on" method, our proposed S2M sampling enables existing unconditional and conditional GANs to draw high-quality multi-label data with a minimal annotation cost. Extensive experiments on real image datasets (*e.g.*, CIFAR-10 and CelebA) verify the effectiveness and correctness of our method, even when compared to a model trained with fully annotated datasets.

1 INTRODUCTION

Since proposed by (Goodfellow et al., 2014), generative adversarial networks (GANs) gained much attention due to its realistic output in a wide range of applications, *e.g.*, image synthesis (Brock et al., 2019; Karras et al., 2018; Park et al., 2019), image translation (Liu et al., 2017; Zhu et al., 2017; Isola et al., 2017a; Choi et al., 2018), and data augmentation (Shrivastava et al., 2016; Bowles et al., 2018). As an advanced task, generating images from given conditions has been achieved by conditional GANs (cGANs) and its variants (Mirza & Osindero, 2014; Odena et al., 2017). Recently, multi-label datasets such as CelebA (Liu et al., 2015) have been introduced in the applications of cGANs (Choi et al., 2018; Lin et al., 2019) to generate diverse images by controlling multiple conditions.

In multi-label data, multiple labels can be assigned to each data instance, where each label is represented as a binary value: 1 for presence and 0 for absence. Nevertheless, the tremendous annotation cost and numerous detection errors still limit the accessibility of multi-label datasets. The trade-off between the quality and the quantity of labels is a common issue for multi-label datasets (Cole et al., 2021). While heuristics for annotations (Lin et al., 2014; Gupta et al., 2019) have been proposed to reduce the cost, they still suffer from detection failures. Under these circumstances, it is natural to consider partially labeled data for the training of GANs.

Consequently, we apply the *single positive setting* (Cole et al., 2021), originally proposed for classification tasks, to conditional generation. In the single positive setting, each data instance is annotated by a *single positive label*; only one positive label is given without any other positive or negative labels. For instance, each facial image in the single positive setting can be labeled as only one of *Black-hair*, *Male*, and *Smile* whereas all attributes are fully specified (*e.g.*, *Smiling black-haired man*) in multi-label datasets. Introducing single positive setting does not only significantly reduce the cost of annotations but also allows the modeling of intrinsic relationships among classes. Here, our aim is to generate data for all possible label combinations that can be appeared in given single positive labels, and this is equivalent to generating multi-label data with at least one positive label.

Recently, several attempts have been made to generate samples of overlapping and non-overlapping classes, which are appeared in the single positive setting. Specially designed generative models (Hou et al., 2018; Asokan & Seelamantula, 2020) have been proposed to exclude samples of a specific

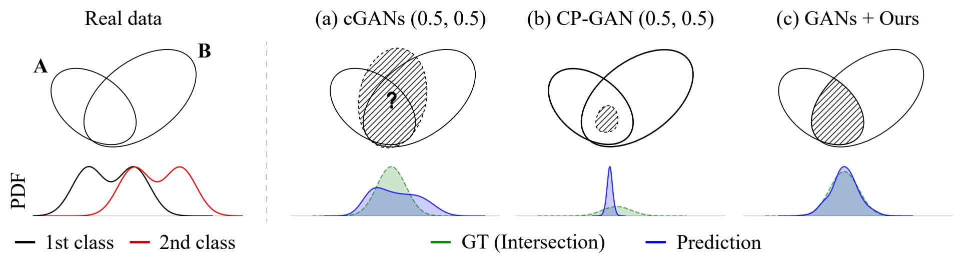


Figure 1: We compare cGANs, CP-GAN and our sampling approach in a class-overlapping case. 1D Gaussian examples consists of two classes of one-dimensional Gaussian mixtures with one common mode, and each method is attempted to generate the overlapping region. For cGANs and CP-GAN, we provide an equal value of 0.5 as labels for two classes. (a) It is not obvious how cGANs obtain samples of the class. (b) CP-GAN draws samples from the narrow region. (c) GANs with S2M sampling can draw samples of the class without sacrificing diversity.

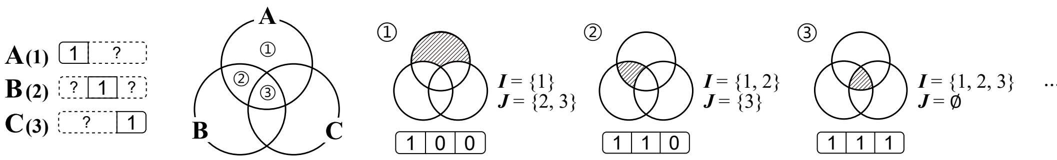
class. Nonetheless, these studies dealt only with two classes and did not consider the generation of samples from overlapping classes. CP-GAN (Kaneko et al., 2019) utilized the probability outputs of the classifier as the labels to capture the relationships between classes. To generate images belonging to n overlapping classes, CP-GAN provides an equal value of $1/n$ as the labels and struggles with generating samples from the true distribution. Figure 1 depicts how different models predict overlapping regions, both conceptually and empirically, with a real 1D Gaussian example.

In this paper, we introduce a novel method called *single-to-multi-label (S2M) Sampling* for generating samples as multi-label data. S2M sampling generates data of all possible combinations of classes only using single positive labels. Concretely, our S2M sampling models overlapping or non-overlapping regions of single positive labels in the sample space of GANs. Using various datasets such as MNIST, FMNIST, CIFAR-10, and CelebA, we show that our S2M sampling correctly samples images as multi-label data in two cases of the single positive setting: (i) datasets of two classes where one class is contained in another class and (ii) datasets of three overlapped classes. Our S2M sampling is designed as a post-processing method attached to pretrained GANs in order to maintain the full generation quality. As a result, our method also can be applied to GANs trained with a large amount of unlabeled data, even with relatively few single positive labels. To the best of our knowledge, our proposed approach is the first sampling algorithm that generates multi-label data while improving the sample quality. Our contributions can be summarized as follows:

- We newly introduce the single positive setting in class-conditional generation and prove that distributions of multi-label data can be derived from that of single positive labels theoretically.
- We propose a novel sampling method called *S2M sampling* to draw samples as multi-label data from GANs, only with single positive labels.
- In diverse settings, we show that the proposed S2M sampling can be applied to various GANs as well as correctly draw samples as multi-label data even in a semi-supervised setting.

2 Related Work

Conditional GANs. The aim of conditional GANs (cGANs) (Mirza & Osindero, 2014) is to model complex distributions and to control data generation by reflecting the label input. Various studies of cGANs have made significant advances in class-conditional image generation by introducing the auxiliary classifier (Odena et al., 2017; Gong et al., 2019), modifying the architecture (Miyato & Koyama, 2018; Brock et al., 2019), and applying metric learning (Kang & Park, 2020). Recently, cGANs have been applied to diverse generation tasks such as image translation (Isola et al., 2017b; Zhu et al., 2017; Choi et al., 2018) and text-to-image generation (Reed et al., 2016; Zhang et al., 2017). In a weakly-supervised setting, GenPU (Hou et al., 2018) and RumiLSGAN (Asokan & Seelamantula, 2020) specify only two classes and draw samples that belong to one class but not the other. CP-GAN (Kaneko et al., 2019) learns to draw samples conditioned on the probability output of the classifier. Given that this model tends to draw samples on a limited region of the data space, it is challenging to ensure the variety of samples as shown in Figure 1. In contrast, we propose a sampling method that draws samples as multi-label data without sacrificing diversity.



(a) Input (b) Overlapping state (c) Possible outputs of S2M sampling

Figure 2: (a) A dataset with single positive labels is given. (b) The regions for each class of A, B, and C overlap in the data space. (c) S2M sampling can draw samples as multi-label data with two index sets of intersection (I) and difference (J).

Sampling in GANs. Sampling methods are used to improve the sample quality in GANs. Discriminator Rejection Sampling (DRS) (Azadi et al., 2019) uses of the scheme of rejection sampling and takes samples close to real data by estimating the density ratio with the discriminator. In addition, Metropolis-Hastings GAN (Turner et al., 2019) adopts Markov chain Monte Carlo (MCMC) method and calibrates the discriminator to improve the sample quality in a high-dimensional data space. Discriminator optimal transport (Tanaka, 2019) utilizes optimal transport theory to obtain realistic samples. Discriminator Driven Latent Sampling (DDLS) (Che et al., 2020) uses the MCMC method in the latent space of GANs to draw realistic samples efficiently. GOLD estimator (Mo et al., 2019) and conditional DDLS (Mo et al., 2019) use sampling algorithms to improve the quality of images for class-conditional generation. While previous studies focus on improving the sample quality, our S2M sampling aims to draw samples as multi-label data while also improving the quality.

3 Mining Multi-Label Samples from Single Positive Labels

3.1 Problem Setting

Let $x \in X$ be a data point as a random variable and let $y_{1:n} \in \{0,1\}^n$ denote its corresponding multi-labels as binary random variables. Here, for every k, $y_k = 1$ indicates that x is contained in the k-th class while $y_k = 0$ indicates that x is not. We consider two index sets, an intersection index set I and a difference index set J, so that the pair (I, J) can be used as an index to indicate the probability density of data points contained in all classes indicated by I but excluded from all classes indicated by J. Let $\mathcal{I}$ be a collection of all possible pairs of I and J, defined as

$$\mathcal{I} = \{(I, J) \in \mathcal{P}(N) \times \mathcal{P}(N) : I \neq \emptyset, I \cap J = \emptyset\}, \tag{1}$$

where $N = \{1, 2, ..., n\}$ is a finite index set of all classes and $\mathcal{P}(N)$ is the power set of N. That is, the intersection index set indicates at least one class and is distinct from the difference index set.

Let $p(x, y_1, y_2, ..., y_n)$ be the joint probability density function, and let $p_{data}(x, c)$ be the joint density of an observable data point $x \in X$ and a class label $c \in N$ such that $p_{data}(x|c) = p(x|y_c = 1)$. Given the class priors $p_{data}(c), \pi_c = p(y_c = 1)$ and samples drawn from the class-conditional density $p_{data}(x|c)$ for each $c = 1, 2, ..., n$, our goal is to draw samples from the conditional density

$$p_{(I,J)}(x) = p(x|\forall i \in I, \forall j \in J, y_i = 1, y_j = 0), \tag{2}$$

for $(I, J) \in \mathcal{I}$, and $\pi_{(I,J)} = p(\forall i \in I, \forall j \in J, y_i = 1, y_j = 0) > 0$.

In this work, we propose to impose a mild constraint which will allow our sampling algorithm to derive the target density, called distinct class separability.

$$\begin{aligned} &\forall i, j \in N \text{ s.t. } i \neq j,\ p(y_i = 1, y_j = 0) > 0 \wedge p(y_j = 1, y_i = 0) > 0 \\ &\qquad \Rightarrow \operatorname{supp} p(x|y_i = 1, y_j = 0) \cap \operatorname{supp} p(x|y_j = 1, y_i = 0) = \emptyset. \end{aligned} \tag{3}$$

This condition states that no data points are likely to be assigned two mutually exclusive labels, which can be naturally assumed in many practical situations. Figure 2 illustrates our problem setting.

3.2 Mining Multi-Label Data with S2M Sampling

A natural question may arise as to whether the supervision given to us is sufficient to obtain samples as multi-label data. To gain insight into this, we initially introduce a useful theorem which provides an alternative formulation for the target density (2).

Theorem 1. *Let $\{f_{(I,J)} : X \to [0,\infty)\}_{(I,J)\in\mathcal{I}}$ be an indexed family of non-negative measurable functions on X, and let $f_k \coloneqq f_{(\{k\},\emptyset)}$. Then, the following conditions hold:*

(a) $\forall (I,J) \in \mathcal{I}, f_{(I,J)} = \sum_{S:I\subseteq S, J\subseteq N\setminus S} f_{(S,N\setminus S)}$

(b) $\forall i,j \in N$ *s.t.* $i \neq j, \operatorname{supp} f_{(\{i\},\{j\})} \cap \operatorname{supp} f_{(\{j\},\{i\})} = \emptyset$

if and only if, for every $(I,J) \in \mathcal{I}$,

$$f_{(I,J)} = \begin{cases} (\min_{i\in I} f_i - \max_{j\in J} f_j)^+ & \textit{if } J \neq \emptyset \\ \min_{i\in I} f_i & \textit{otherwise} \end{cases}, \tag{4}$$

where $(\cdot)^+$ *represents the positive part.*

Proof. Please see Appendix A. □

Let $f_{(I,J)}(x) = p(x, \forall i \in I, \forall j \in J, y_i = 1, y_j = 0)$ for every $(I,J) \in \mathcal{I}$ and assume distinct class separability (3). Then, (a) and (b) in Theorem 1 hold. According to the Theorem 1, if $\pi_{(I,J)} > 0$,

$$p_{(I,J)}(x) = \pi_{(I,J)}^{-1} \left(\min\{\pi_i p(x|y_i = 1) : i \in I\} - \max\{\pi_j p(x|y_j = 1) : j \in J\} \cup \{0\}\right)^+. \tag{5}$$

The alternative formula (5) shows that the target density can be derived from the class-conditional densities of single positive labels. Despite their clear relationships, the conditional density cannot be readily derived during the training procedure; thus, generating images from the target distribution is difficult. In addition, the adjustments for I, J, and the class priors should be allowed in the inference time. To address these issues, instead of training the generative models to model the target distribution directly, we propose the use of our S2M sampling as an "add-on" module to existing generative models. The rest of this section describe the main approach of our S2M sampling.

Density Ratio Estimation. Classification networks are used to compute the ratio between implicitly defined densities of real and fake samples in the literature on GAN sampling (Azadi et al., 2019; Turner et al., 2019; Che et al., 2020). In this work, we utilize this technique to not only compute the density ratio between real and fake samples but also the density ratio between real samples and samples of each single positive class. For simplicity, we denote G as a pretrained generator for both unconditional and conditional GANs. G produces data x by taking a latent z and a class label c for class-conditional generation and only z for unconditional generation. We consider three classifiers D_v, D_r, and D_f which are obtained by minimizing $\mathcal{L}_v$, $\mathcal{L}_r$, and $\mathcal{L}_f$, respectively, *i.e.*,

$$\begin{aligned} \mathcal{L}_v &= -\mathbb{E}_{(x,c)\sim p_{data}(x,c)}[\log D_v(x)] - \mathbb{E}_{x\sim p_G(x)}[\log(1 - D_v(x))] \\ \mathcal{L}_r &= -\mathbb{E}_{(x,c)\sim p_{data}(x,c)}[\log D_r(c|x)], \mathcal{L}_f = -\mathbb{E}_{(x,c)\sim p_G(x,c)}[\log D_f(c|x)]. \end{aligned} \tag{6}$$

The optimal classifiers trained by these losses D_v^*, D_r^*, and D_f^* satisfy the following equations:

$$D_v^*(x) = \frac{p_{data}(x)}{p_{data}(x) + p_G(x)}, D_r^*(c|x) = \frac{p(x|y_c = 1)p_{data}(c)}{p_{data}(x)}, D_f^*(c|x) = \frac{p_G(x|c)p_G(c)}{p_G(x)}. \tag{7}$$

From $D_v^*(x)$, $D_r^*(x)$, and $D_f^*(x)$, we can access the density ratios $p_{data}(x)/p_G(x)$, $p(x|y_c = 1)/p_{data}(x)$, and $p_G(x|c)/p_G(x)$ to compute the acceptance probability of the MCMC method.

S2M Sampling for Unconditional GANs. We apply Metropolis-Hastings (MH) independence sampling (Tierney, 1994; Turner et al., 2019) to draw samples from the complex distribution $p_{(I,J)}$. Assume that the support of p_G contains that of $p_{(I,J)}$. At each step of MH algorithm, we sample a new proposal x' from a proposal distribution $q(x'|x)$ and then accept it with probability $\alpha(x',x) = \min\{1, p_{(I,J)}(x')q(x|x')/p_{(I,J)}(x)q(x'|x)\}$. The chain of samples converges to $p_{(I,J)}$ as MH steps are repeated. We take multiple samples from G as independent proposals, i.e. $q(x'|x) = p_G(x')$, and the acceptance probability $\alpha(x',x)$ is calculated as

$$\begin{aligned} \alpha(x',x) &= \min\left(1, \frac{p_{(I,J)}(x')/p_G(x')}{p_{(I,J)}(x)/p_G(x)}\right) \\ &= \min\left(1, \frac{\left(\min\{r_i(x') : i \in I\} - \max\{r_j(x') : j \in J\} \cup \{0\}\right)^+ \left(D_v^*(x)^{-1} - 1\right)}{\left(\min\{r_i(x) : i \in I\} - \max\{r_j(x) : j \in J\} \cup \{0\}\right)^+ \left(D_v^*(x')^{-1} - 1\right)}\right), \end{aligned} \tag{8}$$

where $r_k(x) := \frac{\pi_k}{p_{data}(k)} D_r^*(k|x)$ for $k \in I \cup J$. To obtain uncorrelated samples, we take a sample after a fixed number of iterations for each chain. We note that the sampling approach allows one to control the parameters I, J, and $\gamma_k = \pi_k / p_{data}(k)$ without any additional training of the model.

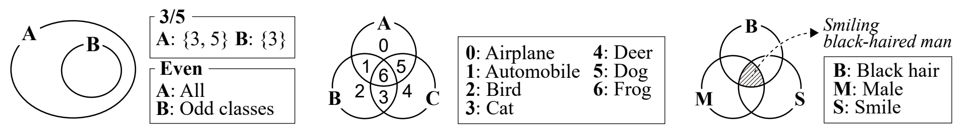
S2M Sampling for Conditional GANs. Conditional GANs can provide a proposal distribution close to the target distribution $p_{(I,J)}$, which greatly increases the sample efficiency of the MCMC method. Let c be a class label such that the support of class-conditional density $p_G(\cdot|c)$ contains that of $p_{(I,J)}$. At each step of the MH algorithm, the proposal $x' \sim q(x'|x) = p_G(x'|c)$ is accepted with a probability $\alpha_c(x', x)$. The desired $\alpha_c(x', x)$ can be calculated as

$$\begin{aligned}\alpha_c(x', x) &= \min\left(1, \frac{p_{(I,J)}(x')/p_G(x'|c)}{p_{(I,J)}(x)/p_G(x|c)}\right) \\ &= \min\left(1, \frac{\left(\min\{r_i(x') : i \in I\} - \max\{r_j(x') : j \in J\} \cup \{0\}\right)^+ D_f^*(c|x)(D_v^*(x)^{-1} - 1)}{\left(\min\{r_i(x) : i \in I\} - \max\{r_j(x) : j \in J\} \cup \{0\}\right)^+ D_f^*(c|x')(D_v^*(x')^{-1} - 1)}\right),\end{aligned} \tag{9}$$

We also apply our sampling approach to another type of a conditional generative model, *e.g.*, CP-GAN. A detailed description of S2M sampling extensions is provided in Appendix B.

Practical Considerations. We employ three classifiers D_v, D_r, and D_f, to compute the acceptance probability used in MCMC method. For better training efficiency of the classifiers, we use shared layers, except for the last linear layer. Since the classifiers are not optimal in practice, we need to calibrate the sampling algorithm. One such approach is temperature scaling (Guo et al., 2017). We adjust the confidence of classification networks by scaling the temperature T. Another approach is to control γ_k. We can obtain confident samples clearly distinct from classes of a difference set by decreasing γ_i for all $i \in I$. We provide the detailed settings in Appendix C, and an ablation study of temperature scaling and the adjustment of γ in Appendix D.

4 Experiments



(a) MNIST / FMNIST (b) CIFAR-10 7to3 (c) CelebA BMS

Figure 3: Experimental settings for each dataset. The corresponding classes$_{\text{orig}}$ are denoted in boxes.

In this section, we validate how our S2M sampling can correctly draw samples as multi-label data from single positive labels. Henceforth, we denote the class defined by the given single positive labels as class$_{\text{single}}$ (*e.g.*, A, B) and each class of the multi-label data target to be derived as class$_{\text{multi}}$ (*e.g.*, $\mathrm{A} \setminus \mathrm{B}$, $\mathrm{A} \cap \mathrm{B}$). In order to avoid ambiguity, the original class is denoted as class$_{\text{orig}}$ (*e.g.*, digits in MNIST or attributes in CelebA). For instance, class$_{\text{single}}$ of even digits in MNIST contains five classes$_{\text{orig}}$: 0, 2, 4, 6, and 8. We conduct the experiments on the synthetic dataset called 2×16 Gaussians and also on image datasets: MNIST, FMNIST, CIFAR-10 and CelebA. 2×16 Gaussians with two overlapped 4×4 grid of Gaussians is designed to verify the correctness of our method. In the image datasets, S2M sampling is examined in two different cases: (i) one class$_{\text{single}}$ contained in another class$_{\text{single}}$, and (ii) three different classes$_{\text{single}}$ that overlap (See Figure 3).

As the choice of base model, we use GANs with fully connected layers for the synthetic data, MNIST, and FMNIST. For CIFAR-10 and CelebA, we choose SNGAN ResNet (Miyato et al., 2018) architecture for stable training. For each dataset, we use the same architecture of GANs except for GenPU, which requires two generators and three discriminators. As the classifiers used for S2M sampling, we use a simple linear classifier for 2×16 Gaussians, LeNet5 (Lecun et al., 1998) for MNIST and FMNIST, and MobileNetV2 (Sandler et al., 2018) for CIFAR-10 and CelebA. All classifiers used for S2M sampling are only trained with the labels of classes$_{\text{single}}$.

For evaluation metrics, we mainly use accuracy, Fréchet Inception Distance (FID) (Heusel et al., 2017), and Inception Score (IS) (Salimans et al., 2016). The accuracy measures how many images are assigned to the target class$_{\text{multi}}$ by a classifier pre-trained with fully annotated labels (*i.e.*, classes$_{\text{orig}}$). For accuracy, we use LeNet5 for MNIST and FMNIST, and MobileNetV2 for CIFAR-10 and CelebA. FID/IS are widely used to evaluate both the quality and diversity of generated images. We use test datasets as reference images to compute FID. Since FID/IS are not applicable for non-real images, *e.g.*, MNIST or FMNIST, we evaluate FID with the activations of pretrained LeNet5 for those datasets and denote these as FID†. We also provide Precision and Recall (Sajjadi et al., 2018) in Appendix D for further analysis. We compute the metrics using $10k$ samples for each class$_{\text{multi}}$. All results of the experiments are averaged over three independent trials, and the standard deviation is denoted by subscripts. More details are provided in Appendix C.

4.1 Two Classes Gaussian Example

5×5 grid of two-dimensional Gaussians called *25 Gaussians* is commonly used to validate the correctness of sampling approaches (Azadi et al., 2019; Turner et al., 2019; Che et al., 2020). In our case, we modify *25 Gaussians* to have two 4×4 grids of two-dimensional Gaussians of overlapped classes A, B, denoted as **$\mathbf{2} \times \mathbf{16}$ Gaussians** (See Figure 4). The modes are horizontally and vertically spaced by 1.0 and have a standard deviation of 0.05. We first train GANs with randomly drawn points within two grids. Then, we adopt S2M sampling to GANs with a classifier trained with points of A and B. We obtain samples at 400 MC iterations. As shown in Figure 4, our S2M sampling accurately estimates various conditional densities ($A, B, A \setminus B, B \setminus A, A \cap B$) while improving the quality of the points. On the other hand, GANs tends to generate spurious lines between the points.

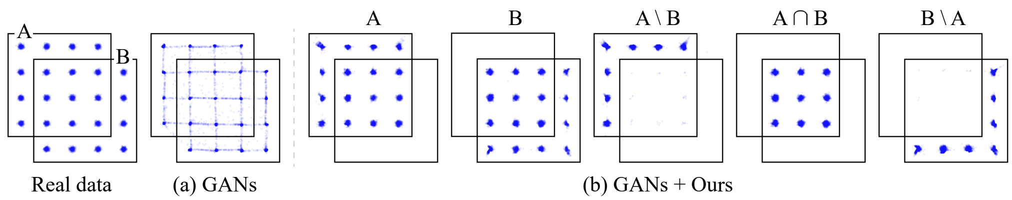


Figure 4: Example of 2×16 Gaussians. Compared to the base model, our S2M Sampling improves the quality of the points while sampling points within various conditions $(\mathrm{A}, \mathrm{B}, \mathrm{A} \setminus \mathrm{B}, \mathrm{B} \setminus \mathrm{A}, \mathrm{A} \cap \mathrm{B})$ accurately even with single positive labels (A, B).

For the quantitative analysis, we report the accuracy, high-quality ratio, and mode standard deviation in Table 1. We generate $10k$ samples and assign each point to the mode with the closest L_2 distance for measuring the accuracy. Following (Turner et al., 2019), samples whose L_2 distances are within four standard deviations are considered as “high-quality” samples. The results indicate that our S2M sampling favorably obtains high-quality samples for various conditions. Apart from accuracy, the ratio of high-quality samples is improved by 14.36% on average.

Table 1: Accuracy (%), high-quality ratio (%), and mode standard deviation on 2×16 Gaussians.

Condition	GANs			**GANs + Ours**		
	Accuracy	High quality	Mode S.D.	Accuracy	High quality	Mode S.D.
A, B	$69.83_{\pm 0.35}$	$84.39_{\pm 0.60}$	$0.106_{\pm 0.002}$	$100.00_{\pm 0.00}$	$98.94_{\pm 0.40}$	$0.052_{\pm 0.002}$
$\mathrm{A} \setminus \mathrm{B}, \mathrm{B} \setminus \mathrm{A}$	$30.17_{\pm 0.35}$	$88.87_{\pm 0.50}$	$0.090_{\pm 0.002}$	$99.52_{\pm 0.36}$	$98.67_{\pm 0.51}$	$0.051_{\pm 0.002}$
$\mathrm{A} \cap \mathrm{B}$	$39.66_{\pm 0.46}$	$80.98_{\pm 0.82}$	$0.118_{\pm 0.003}$	$100.00_{\pm 0.00}$	$99.73_{\pm 0.14}$	$0.050_{\pm 0.001}$

4.2 MNIST & FMNIST: Classes with Inclusion Relationship

In this section, we consider a special case of our problem setting, where one class is contained in another class. This also can be considered as a positive-unlabeled (Denis, 1998; Denis et al., 2005) setting if the smaller class and the remainder represent positive data and unlabeled data, respectively. Under this constraint, GenPU (Hou et al., 2018) and RumiLSGAN (Asokan & Seelamantula, 2020) can be applied to exclude samples in the smaller class. We compare our S2M sampling with GenPU, RumiLSGAN, and CP-GAN (Kaneko et al., 2019) in three settings: **(i) MNIST 3/5** ($\{3, 5\} \setminus \{3\}$), **(ii) MNIST Even** ($\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\} \setminus \{1, 3, 5, 7, 9\}$), and **(iii) FMNIST Even** ($\{0_{\text{T-shirt/Top}}, 1_{\text{Trouser}}, 2_{\text{Pullover}}, 3_{\text{Dress}}, 4_{\text{Coat}}, 5_{\text{Sandal}}, 6_{\text{Shirt}}, 7_{\text{Sneaker}}, 8_{\text{Bag}}, 9_{\text{Ankle boot}}\} \setminus \{1, 3, 5, 7, 9\}$). We apply our S2M sampling to unconditional GANs and sample images at 100 MC iterations.

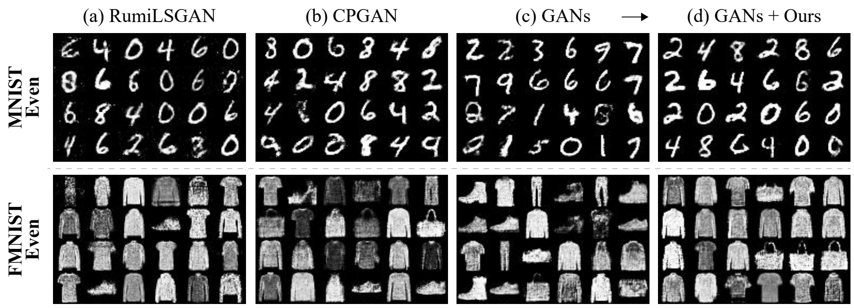


Figure 5: Qualitative results of baselines and S2M sampling for MNIST Even and FMNIST Even. We apply our S2M sampling to unconditional GANs specified in (c).

Table 2: Accuracy (%) and FID† for various models on MNIST Even and FMNIST Even.

Model	MNIST 3/5		MNIST Even		FMNIST Even	
	Acc. (↑)	FID† (↓)	Acc. (↑)	FID† (↓)	Acc. (↑)	FID† (↓)
GenPU	99.18±0.64	1.82±1.12	-	-	-	-
RumiLSGAN	77.92±1.74	12.13±1.13	85.79±4.82	3.45±1.47	90.99±0.84	3.12±0.25
CPGAN	68.23±1.17	18.62±0.82	87.88±0.80	2.17±0.17	81.22±0.53	6.14±0.60
GANs	47.91±0.15	35.28±0.13	46.89±0.65	21.90±0.74	48.16±0.79	28.43±0.28
GANs + Ours	**99.64**±0.25	**0.78**±0.27	**96.35**±0.24	**0.86**±0.21	**97.95**±0.61	**2.44**±0.29

As reported in Table 2, our S2M sampling always outperforms the baselines, even with unconditional GANs. For GenPU, the performance is reported only for MNIST 3/5 due to its mode collapse issue (Chen et al., 2020; Chiaroni et al., 2020). Most notably, our S2M sampling improves the accuracy by 8.47% and 6.96% compared to the second-best models on MNIST Even and FMNIST Even. Figure 5 shows that baselines struggle to eliminate the smaller class$_{\text{single}}$ completely (*e.g.*, odd digits for MNIST or *Sneaker* for FMNIST). In contrast, our S2M sampling added on unconditional GANs samples images within the target class$_{\text{multi}}$ accurately. As mentioned in Section 3, our S2M sampling can be applied to conditional GANs such as CP-GAN, which we will examine later.

4.3 Real data: Classes with Overlapped Regions

In this section, we construct datasets consisting of three overlapped classes$_{\text{single}}$ on CIFAR-10 and CelebA. For CIFAR-10, we use the dataset called **CIFAR-10 7to3** proposed by (Kaneko et al., 2019), where each class$_{\text{single}}$ contains four classes$_{\text{orig}}$ (*e.g.*, $\{$Airplane, Automobile, Dog, Frog$\} \subseteq$ A) so that each class$_{\text{multi}}$ can be seen as a single class$_{\text{orig}}$. In the CelebA case, we assign *Black hair*, *Male*, and *Smile* attributes as a class$_{\text{single}}$ and denote this case as **CelebA BMS**, as shown in Figure 3.

For scrutiny, we introduce two versions of cGANs and ACGAN as additional baselines: models trained with fully specified labels (oracle models) and those with single positive labels (marked with *). The oracle models are the topmost baselines, as these models already have full knowledge of classes$_{\text{multi}}$. cGANs* and ACGAN* are introduced to demonstrate how the existing models behave in the single positive setting. To predict each class$_{\text{multi}}$, the exact label of the class$_{\text{multi}}$ is simply given for the oracle models, and labels of $1/m$ values are given to each class$_{\text{single}}$ to generate samples within the intersection of m classes$_{\text{single}}$ for cGAN* and ACGAN* as introduced in (Kaneko et al., 2019). For evaluation, we assess seven classes$_{\text{multi}}$ derived by three overlapped classes$_{\text{single}}$ as illustrated in the first row of Figure 6. In all experiments, only single positive labels are used except for the oracle models, and samples are obtained at 200 MC iterations for the sampling method. Our S2M sampling is adopted on unconditional GANs, cGANs and CP-GAN.

Table 3 shows the average results of seven classes$_{\text{multi}}$ for different models. Regardless of the dataset, our S2M sampling always proves superior to the base models. Compared to the best baseline, our S2M sampling improves the accuracy by 25.42% for CIFAR-10 7to3 and 15.84% for CelebA BMS, while decreasing FID. More surprisingly, even with single positive labels, our S2M sampling shows comparable performance to the oracle models. The visual results of our proposed S2M sampling are depicted in Figure 6 (See Appendix D for more results). The results indicate that our S2M sampling precisely estimates the true distributions for both overlapping and non-overlapping classes.

Table 3: Accuracy (%), FID and IS for different models on CIFAR-10 7to3 and CelebA BMS.

Model	CIFAR-10 7to3			CelebA BMS		
	Acc. (↑)	FID (↓)	IS (↑)	Acc. (↑)	FID (↓)	IS (↑)
cGANs (Oracle)	86.47±0.10	15.07±0.41	8.40±0.09	68.31±1.32	9.71±0.19	2.47±0.03
ACGAN (Oracle)	90.08±0.95	15.63±0.44	8.28±0.10	72.98±0.40	7.82±0.05	2.46±0.02
cGANs*	25.37±0.29	22.37±0.26	7.61±0.07	27.59±0.49	9.69±0.63	2.50±0.01
ACGAN*	29.49±0.95	23.44±0.44	7.58±0.10	29.15±0.40	10.83±0.05	2.32±0.02
CPGAN	65.30±0.56	24.11±1.03	7.66±0.04	58.70±4.56	21.98±0.65	2.32±0.04
GANs + Ours	80.30±1.02	16.79±0.25	8.04±0.04	73.92±3.01	**8.53±0.86**	**2.53±0.02**
cGANs + Ours	84.44±0.66	**16.36±0.25**	**8.28±0.06**	**74.54±2.79**	9.76±0.74	**2.53±0.01**
CPGAN + Ours	**90.72±1.33**	22.54±0.94	7.79±0.04	72.99±0.66	21.30±0.62	2.25±0.02

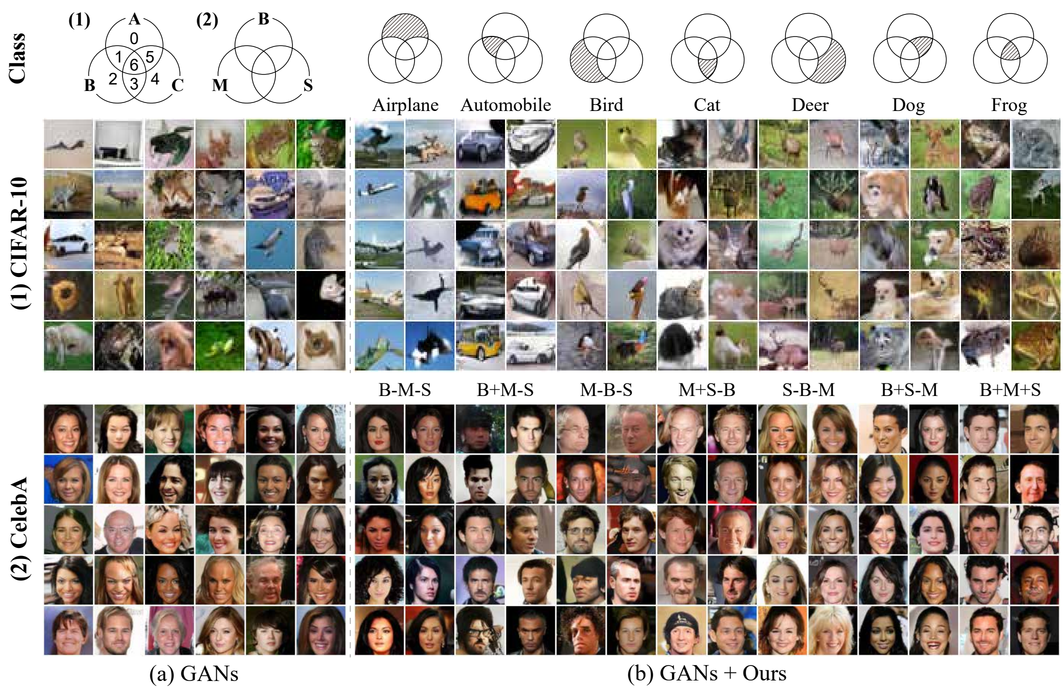


Figure 6: Results of our S2M sampling with unconditional GANs on CIFAR-10 7to3 and CelebA BMS. The first row depicts the target class$_{\text{multi}}$. Intersections and differences are denoted by plus signs and minus signs, respectively.

5 Discussion

5.1 Improving Sample Efficiency with Latent Candidates

While our S2M sampling allows one to draw samples as multi-label data, multiple samples should be rejected in the sampling procedure. Specifically, many samples can be wasted if the samples in the target class$_{\text{multi}}$ rarely appear in the original dataset. To tackle this issue, we propose a simple heuristic algorithm for selecting “latent candidates”. We hypothesize that the latent samples of GANs corresponding to the target distribution tend to be close to each other in the latent space. We initially draw a certain number of pilot samples $x_{1:m}$ using S2M sampling, and then obtain the corresponding latent samples $z_{1:m}$; $x_k = G(z_k)$. Since the latent samples are nearly restricted to the latent prior of GANs, we can derive an approximated distribution $\hat{p}_{z_t}$ of the target latent samples by fitting a simple probabilistic model such as Gaussian Mixture Model (GMM) using $z_{1:m}$. By taking latents from $\hat{p}_{z_t}$, we can further improve the sample efficiency. For evaluation, we examine GMM on two classes$_{\text{multi}}$ with low acceptance probabilities: Frog in CIFAR-10 7to3 and B+M+S in CelebA BMS. For each class$_{\text{multi}}$, we fit GMM of eight components with expectation–maximization algorithm (Bishop, 2007) using $10k$ pilot samples obtained at 100 MC iterations.

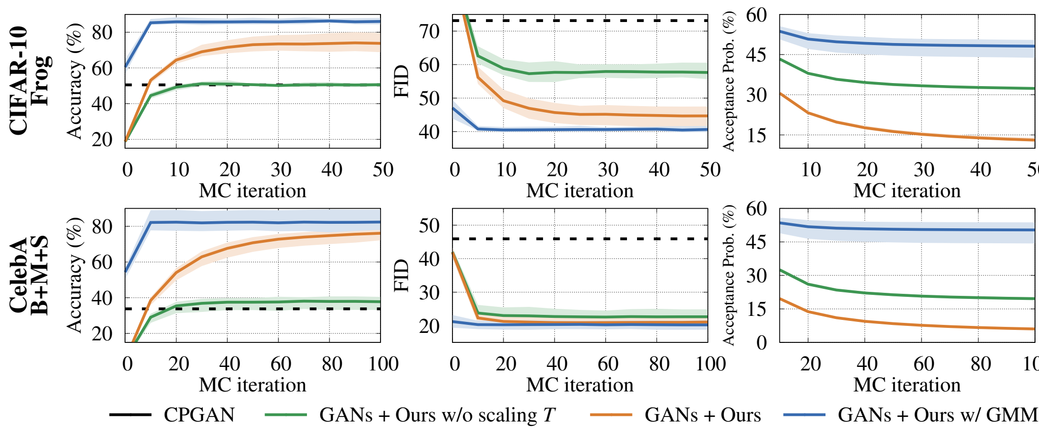


Figure 7: Results per MC iteration for our S2M sampling with and without GMM.

Figure 7 shows the accuracy, FID, and the acceptance probability per MC iteration. As when temperature scaling is not performed (green line), a high acceptance probability is commonly caused by an incorrectly predicted distribution. However, GMM greatly improves the acceptance probability from the beginning without an accuracy drop. For future work, a more efficient sampling such as adaptive MCMC (Gilks & Wild, 1992; Gilks et al., 1995) may further improve our S2M sampling.

5.2 Mining Samples in Semi-Supervised Setting

Single positive labels require much lower cost than multi-label datasets. Nevertheless, large amounts of unlabeled data are most commonly provided in the real world. In this section, we explore a semi-supervised setting in which, only small amounts of a dataset have single positive labels. Ideally, if the given labeled data is sufficient for training classifiers, we can accurately sample the images while fully utilizing the unlabeled data for GANs. This assumption is presumable, as the discriminator is known to reach the optimal easily compared to the generator (Turner et al., 2019). Concretely, we conduct the experiments of CelebA BMS as in Section 4.3 but with less labeled data. With unconditional GANs, we examine our S2M sampling with various ratios of single positive labels: 50%, 20%, and 10% of the dataset. As shown in Table 4, even when only 10% of the labels are given, S2M sampling still outperforms the baselines in Table 3 with only an approximate 4% accuracy drop. The results show that our S2M sampling can be used in a semi-supervised setting for sampling high quality images within the desired class$_{\text{multi}}$ without degrading the generation capability of GANs.

Table 4: Accuracy (%), FID, and IS in CelebA BMS for various levels of supervision.

Supervision ratio	All (100%)	50%	20%	10%
Acc. (↑)	73.92$\pm$3.01	72.91$\pm$2.24	71.46$\pm$2.62	70.01$\pm$3.35
FID (↓)	8.53$\pm$0.86	8.61$\pm$1.03	8.71$\pm$1.02	8.79$\pm$0.77
IS (↑)	2.53$\pm$0.02	2.55$\pm$0.05	2.52$\pm$0.04	2.49$\pm$0.01

6 Conclusion

We investigate the single positive setting in the class-conditional generation task and propose a novel method called S2M sampling for drawing samples as multi-label data only from single positive labels. We demonstrate that our proposed S2M sampling can be adopted to a variety of GANs and accurately draw samples as multi-label data with a minimal annotation cost. Moreover, we introduce GMM as a simple yet effective method to improve the sampling efficiency and show that our S2M sampling remains effective in a semi-supervised setting. We hope that existing models can be further improved with the augmented multi-label dataset obtained by our S2M sampling. In future works, S2M sampling can be employed to datasets that have much more complex relationships among classes or that contain classes with imbalances.

7 Ethics Statement

This work demonstrates that it is possible to generate multi-label data from limited labels. In addition, this work can be freely adopted to unconditional GANs trained with a large amount of unlabeled data. Hence, our work can reduce the high annotation cost that research groups face in common. Despite the fact that deep learning models tend to struggle from learning underrepresented data (Mehrabi et al., 2019), properly calibrated sampling algorithm does not readily ignore rarely appearing data, meaning that it is unlikely to introduce bias into generative models.

8 Reproducibility Statement

For reproducibility, we provide executable source codes of our proposed S2M sampling with instructions in the Supplementary Material.

References

Siddarth Asokan and Chandra Sekhar Seelamantula. Teaching a GAN what not to learn. In *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Samaneh Azadi, Catherine Olsson, Trevor Darrell, Ian J. Goodfellow, and Augustus Odena. Discriminator rejection sampling. In *Proc. the International Conference on Learning Representations (ICLR)*, 2019.

Christopher M. Bishop. *Pattern recognition and machine learning, 5th Edition*. 2007.

Christopher Bowles, Liang Chen, Ricardo Guerrero, Paul Bentley, Roger N. Gunn, Alexander Hammers, David Alexander Dickie, Maria del C. Valdés Hernández, Joanna M. Wardlaw, and Daniel Rueckert. GAN augmentation: Augmenting training data using generative adversarial networks. *CoRR*, abs/1810.10863, 2018.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *Proc. the International Conference on Learning Representations (ICLR)*, 2019.

Tong Che, Ruixiang Zhang, Jascha Sohl-Dickstein, Hugo Larochelle, Liam Paull, Yuan Cao, and Yoshua Bengio. Your GAN is secretly an energy-based model and you should use discriminator driven latent sampling. In *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Hui Chen, Fangqing Liu, Yin Wang, Liyue Zhao, and Hao Wu. A variational approach for learning from positive and unlabeled data. In *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Florent Chiaroni, Ghazaleh Khodabandelou, Mohamed-Cherif Rahal, Nicolas Hueber, and Frédéric Dufaux. Counter-examples generation from a positive unlabeled image dataset. *Pattern Recognition*, 107:107527, 2020.

Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018.

Elijah Cole, Oisin Mac Aodha, Titouan Lorieul, Pietro Perona, Dan Morris, and Nebojsa Jojic. Multi-label learning from single positive labels. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2021.

Harm de Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C. Courville. Modulating early visual processing by language. In *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

François Denis. PAC learning from positive statistical queries. In *ALT*, Lecture Notes in Computer Science, 1998.

François Denis, Rémi Gilleron, and Fabien Letouzey. Learning from positive and unlabeled examples. *Theor. Comput. Sci.*, 2005.

W. R. Gilks and P. Wild. Adaptive rejection sampling for gibbs sampling. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 41(2):337–348, 1992.

W. R. Gilks, N. G. Best, and K. K. C. Tan. Adaptive rejection metropolis sampling within gibbs sampling. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 44(4):455–472, 1995.

Mingming Gong, Yanwu Xu, Chunyuan Li, Kun Zhang, and Kayhan Batmanghelich. Twin auxiliary classifiers GAN. *CoRR*, abs/1907.02690, 2019.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proc. the International Conference on Machine Learning (ICML)*, 2017.

Agrim Gupta, Piotr Dollár, and Ross B. Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2019.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems*, pp. 6626–6637, 2017.

Ming Hou, Brahim Chaib-draa, Chao Li, and Qibin Zhao. Generative adversarial positive-unlabelled learning. In *Proc. the International Joint Conference on Artificial Intelligence (IJCAI)*, 2018.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2017a.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2017b.

Takuhiro Kaneko, Yoshitaka Ushiku, and Tatsuya Harada. Class-distinct and class-mutual image generation with gans. In *Proc. of the British Machine Vision Conference (BMVC)*, 2019.

Minguk Kang and Jaesik Park. Contragan: Contrastive learning for conditional image generation. In *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *Proc. the International Conference on Learning Representations (ICLR)*, 2018.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. the International Conference on Learning Representations (ICLR)*, 2015.

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

Jae Hyun Lim and Jong Chul Ye. Geometric GAN. *CoRR*, abs/1705.02894, 2017.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *Proc. of the European Conference on Computer Vision (ECCV)*, 2014.

Yu-Jing Lin, Po-Wei Wu, Che-Han Chang, Edward Y. Chang, and Shih-Wei Liao. Relgan: Multi-domain image-to-image translation via relative attributes. In *Proc. of the IEEE international conference on computer vision (ICCV)*, 2019.

Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proc. of the IEEE international conference on computer vision (ICCV)*, 2015.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *CoRR*, abs/1908.09635, 2019.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014.

Takeru Miyato and Masanori Koyama. cgans with projection discriminator. In *Proc. the International Conference on Learning Representations (ICLR)*, 2018.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *Proc. the International Conference on Learning Representations (ICLR)*, 2018.

Sangwoo Mo, Chiheon Kim, Sungwoong Kim, Minsu Cho, and Jinwoo Shin. Mining GOLD samples for conditional gans. In *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *Proc. the International Conference on Machine Learning (ICML)*, 2017.

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2019.

Scott E. Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *Proc. the International Conference on Machine Learning (ICML)*, 2016.

Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018.

Abhinav Shrivastava, Abhinav Gupta, and Ross B. Girshick. Training region-based object detectors with online hard example mining. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016.

Akinori Tanaka. Discriminator optimal transport. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Luke Tierney. Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4): 1701–1728, 1994.

Ryan D. Turner, Jane Hung, Eric Frank, Yunus Saatchi, and Jason Yosinski. Metropolis-hastings generative adversarial networks. In *Proc. the International Conference on Machine Learning (ICML)*, 2019.

Xiang Wei, Boqing Gong, Zixia Liu, Wei Lu, and Liqiang Wang. Improving the improved training of wasserstein gans: A consistency term and its dual effect. In *Proc. the International Conference on Learning Representations (ICLR)*, 2018.

Han Zhang, Tao Xu, and Hongsheng Li. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proc. of the IEEE international conference on computer vision (ICCV)*, 2017.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. of the IEEE international conference on computer vision (ICCV)*, 2017.

A Proof of main theorem

Lemma 1. *Assume that (a) and (b) in Theorem 1 hold.* $\forall v \in N, \forall U \subseteq N$ *s.t.* $v \notin U \neq \emptyset$, $\operatorname{supp} f_{(U,\{v\})} \cap \operatorname{supp} f_{(\{v\},U)} = \emptyset$.

Proof. Choose any $u \in U$. Then,

$$\begin{aligned} f_{(\{u\},\{v\})} &= \left(\sum\nolimits_{S: u\in S, v\in N\setminus S, U\not\subseteq S} f_{(S,N\setminus S)}\right) + f_{(U,\{v\})} \geq f_{(U,\{v\})} \\ f_{(\{v\},\{u\})} &= \left(\sum\nolimits_{S: u\in S, v\in N\setminus S, U\not\subseteq N\setminus S} f_{(S,N\setminus S)}\right) + f_{(\{v\},U)} \geq f_{(\{v\},U)}. \end{aligned} \tag{10}$$

Therefore, $\operatorname{supp} f_{(U,\{v\})} \cap \operatorname{supp} f_{(\{v\},U)} \subseteq \operatorname{supp} f_{(\{u\},\{v\})} \cap \operatorname{supp} f_{(\{v\},\{u\})}$ which implies $\operatorname{supp} f_{(U,\{v\})} \cap \operatorname{supp} f_{(\{v\},U)} = \emptyset$. □

Lemma 2. *Assume that (a) and (b) in Theorem 1 hold.* $\forall I \subseteq N$ *s.t.* $I \neq \emptyset$, $f_{(I,\emptyset)} = \min_{i\in I} f_i$.

Proof. We will use induction to prove the lemma. Let $P(k)$ be the following statement.

$$P(k): \forall I \text{ s.t. } 1 \leq |I| = k \leq |N|, \text{then } f_{(I,\emptyset)} = \min_{i\in I} f_i. \tag{11}$$

For the base case $k = 1$, the statement holds by the definition. Assume that the induction hypothesis for $k \leq l < |N|$ holds. Consider $|I| = l + 1$ and choose any $i \in I$. Then,

$$\begin{aligned} \min_{i\in I} f_i &= \min\{f_{(I\setminus\{i\},\emptyset)}, f_{(\{i\},\emptyset)}\} && \text{By the induction hypothesis} \\ &= f_{(\{i\},\emptyset)} - \max\{f_{(\{i\},\emptyset)} - f_{(I\setminus\{i\},\emptyset)}, 0\} \\ &= f_{(\{i\},\emptyset)} - \max\{f_{(\{i\},I\setminus\{i\})} - f_{(I\setminus\{i\},\{i\})}, 0\} \\ &= f_{(\{i\},\emptyset)} - f_{(\{i\},I\setminus\{i\})} && \text{By Lemma1} \\ &= f_{(I,\emptyset)} \end{aligned} \tag{12}$$

Therefore, $P(l + 1)$ holds. We conclude that $f_{(I,\emptyset)} = \min_{i\in I} f_i$ for $\emptyset \neq I \subseteq N$. □

Theorem 1. *Let* $\{f_{(I,J)} : X \to [0,\infty)\}_{(I,J)\in\mathcal{I}}$ *be an indexed family of non-negative measurable functions on* X*, and let* $f_k \coloneqq f_{(\{k\},\emptyset)}$*. Then, the following conditions hold:*

(a) $\forall (I, J) \in \mathcal{I}, f_{(I,J)} = \sum_{S: I\subseteq S, J\subseteq N\setminus S} f_{(S,N\setminus S)}$

(b) $\forall i, j \in N$ *s.t.* $i \neq j, \operatorname{supp} f_{(\{i\},\{j\})} \cap \operatorname{supp} f_{(\{j\},\{i\})} = \emptyset$

if and only if, for every $(I, J) \in \mathcal{I}$,

$$f_{(I,J)} = \begin{cases} (\min_{i\in I} f_i - \max_{j\in J} f_j)^+ & \text{if } J \neq \emptyset \\ \min_{i\in I} f_i & \text{otherwise} \end{cases}, \tag{4}$$

where $(\cdot)^+$ *represents the positive part.*

Proof. We first show the necessity of the condition. Assume that (a) and (b) hold. If $J = \emptyset$, then $f_{(I,J)} = \min_{i\in I} f_i$ by Lemma2. Hence, we may assume that $J \neq \emptyset$. Fix $x \in X$, and let $\{a_1, a_2, ..., a_{|J|}\}$ be an arrangement of J so that $f_{a_i}(x) \leq f_{a_j}(x)$ for all $i < j$. For every $\emptyset \neq S \subseteq J$, we let $m(S)$ denote the minimum index s such that $a_s \in S$.

Note that

$$\begin{aligned} f_{(I,J)}(x) &= \sum_{S\subseteq J} (-1)^{|S|} f_{(I\cup S,\emptyset)}(x) && \text{By Inclusion–exclusion principle} \\ &= \sum_{S\subseteq J} (-1)^{|S|} \min_{i\in I\cup S} f_i(x) && \text{By Lemma2} \end{aligned} \tag{13}$$

We now decompose the last summation into three cases.

(i) $S = \emptyset$

$$(-1)^{|S|} \min_{i \in I \cup S} f_i(x) = \min_{i \in I} f_i(x). \tag{14}$$

(ii) $m(S) < |J|$

$$\begin{aligned}\sum_{S:m(S)<|J|} (-1)^{|S|} \min_{i \in I \cup S} f_i(x) &= \sum_{j<|J|} \sum_{S:m(S)=j} (-1)^{|S|} \min_{i \in I \cup \{a_j\}} f_i(x) \\ &= \sum_{j<|J|} \left(\min_{i \in I \cup \{a_j\}} f_i(x) \right) \left\{ (-1) \cdot 2^{|J|-j-1} + 2^{|J|-j-1} \right\} \\ &= 0.\end{aligned} \tag{15}$$

(iii) $m(S) = |J|$

$$(-1)^{|S|} \min_{i \in I \cup S} f_i(x) = - \min_{i \in I \cup \{a_{|J|}\}} f_i(x). \tag{16}$$

Summing up all of the above terms gives the rest result.

$$\begin{aligned}f_{(I,J)}(x) &= \min_{i \in I} f_i(x) - \min_{i \in I \cup \{a_{|J|}\}} f_i(x) \\ &= \min_{i \in I} f_i(x) - \min\{\min_{i \in I} f_i(x), \max_{j \in J} f_j(x)\} \\ &= \left(\min_{i \in I} f_i(x) - \max_{j \in J} f_j(x) \right)^+ .\end{aligned} \tag{17}$$

To show the sufficiency, assume

$$\forall (I,J) \in \mathcal{I}, f_{(I,J)} = \begin{cases} (\min_{i \in I} f_i - \max_{j \in J} f_j)^+ & \text{if } J \neq \emptyset \\ \min_{i \in I} f_i & \text{otherwise} \end{cases}. \tag{18}$$

Let us assume that $f_{(I,J)} \neq \sum_{S:I \subseteq S, J \subseteq N \setminus S} f_{(S,N \setminus S)}$ for some $(I,J) \in \mathcal{I}$. Choose such I, J so that $|I| + |J|$ is maximum. Note that $I \cup J \subsetneq N$ because $\sum_{S:I \subseteq S, J \subseteq N \setminus S} f_{(S,N \setminus S)}$ is exactly the same expression as $f_{(I,J)}$ for $I \cup J = N$. Hence, we can choose some $k \in N \setminus (I \cup J)$. By the maximality of $|I| + |J|$, the following two equations hold.

$$\begin{aligned}f_{(I,J \cup \{k\})} &= \sum\nolimits_{S:I \subseteq S, J \cup \{k\} \subseteq N \setminus S} f_{(S,N \setminus S)} \\ f_{(I \cup \{k\},J)} &= \sum\nolimits_{S:I \cup \{k\} \subseteq S, J \subseteq N \setminus S} f_{(S,N \setminus S)}.\end{aligned} \tag{19}$$

We use above two equations and consider all possible inequalities among $\min_{i \in I} f_i$, $max_{j \in J} f_j$, and f_k. The following equation always holds regardless of these inequalities.

$$\begin{aligned}\sum_{S:I \subseteq S, J \subseteq N \setminus S} f_{(S,N \setminus S)} &= f_{(I,J \cup \{k\})} + f_{(I \cup \{k\},J)} \\ &= \begin{cases} \left(\min_{i \in I} f_i - \max_{j \in J \cup \{k\}} f_j\right)^+ + \left(\min_{i \in I \cup \{k\}} f_i - \max_{j \in J} f_j\right)^+ & \text{if } J \neq \emptyset \\ (\min_{i \in I} f_i - f_k)^+ + \min_{i \in I \cup \{k\}} f_i & \text{otherwise} \end{cases} \\ &= \begin{cases} (\min_{i \in I} f_i - \max_{j \in J} f_j)^+ & \text{if } J \neq \emptyset \\ \min_{i \in I} f_i & \text{otherwise} \end{cases} \\ &= f_{(I,J)},\end{aligned} \tag{20}$$

which leads to a contradiction.

Also, for every $i, j \in N$ such that $i \neq j$,

$$\begin{aligned}\min(f_{(\{i\},\{j\})}, f_{(\{j\},\{i\})}) &= \min\{(f_i - f_j)^+, (f_j - f_i)^+\} \\ &= (\min\{f_i - f_j, f_j - f_i\})^+ \\ &= 0.\end{aligned} \tag{21}$$

Therefore, (a) and (b) hold. $\square$

B Description for S2M Sampling

In Section 3.2, we describe how to build S2M sampling upon unconditional GANs, and class-conditional GANs. We also adopt our S2M sampling to another type of conditional generative model, *e.g.*CP-GAN (Kaneko et al., 2019). In this case, the generator G takes $w(x)$ as a condition where w is a deterministic function. Let c be a condition such that the support of conditional density $p_G(\cdot|c)$ contains that of $p_{(I,J)}$. At each step of MH algorithm, the proposal $x' \sim q(x'|x) = p_G(x'|c)$ is accepted with a probability $\alpha_w(x', x)$. If we assume $w(x') = w(x) = c$, then $p_G(x'|c)/p_G(x|c) = p_G(x')/p_G(x)$. Hence, the desired $\alpha_w(x', x)$ can be calculated similarly to $\alpha(x', x)$:

$$\begin{aligned}\alpha_w(x', x) &= \min\left(1, \frac{p_{(I,J)}(x')/p_G(x'|c)}{p_{(I,J)}(x)/p_G(x|c)}\right) \\ &= \min\left(1, \frac{\left(\min\{r_i(x') : i \in I\} - \max\{r_j(x') : j \in J\} \cup \{0\}\right)^+ \left(D_v^*(x)^{-1} - 1\right)}{\left(\min\{r_i(x) : i \in I\} - \max\{r_j(x) : j \in J\} \cup \{0\}\right)^+ \left(D_v^*(x')^{-1} - 1\right)}\right),\end{aligned} \tag{22}$$

Although the support of conditional density of CP-GAN may not cover all that of $p_{(I,J)}$, S2M sampling can still be used to draw confident samples suitable for a given condition. We empirically show that our S2M sampling algorithm can improve the generation accuracy while maintaining almost the diversity of samples (Section 4.3).

Algorithm 1 illustrates the use of S2M sampling for GANs. This algorithm can be easily extended to the conditional versions, *e.g.*, cGANs and CP-GAN, by replacing the acceptance probability (See Equation 9 and Equation 22).

Algorithm 1 *S2M Sampling for GANs*

Input: generator G, classifiers D_v^*, D_r^*, intersection index set I, difference index set J, and class prior ratios $\gamma_{1:N}$

Output: filtered sample x

```
Draw x from G such that x ∈ supp p_(I,J).
for k = 1 to K do
    Draw x' from G.
    Draw u from Uniform(0,1).
    r_i ← γ_i D_r^*(i|x) for every i ∈ I ∪ J
    r'_i ← γ_i D_r^*(i|x') for every i ∈ I ∪ J
    α ← min(1, (min{r'_i : i∈I} − max{r'_j : j∈J} ∪ {0})^+ (D_v^*(x)^{-1} − 1) / (min{r_i : i∈I} − max{r_j : j∈J} ∪ {0})^+ (D_v^*(x')^{-1} − 1))    ▷ acceptance probability
    if u ≤ α then
        x ← x'
    end if
end for
```

C Experiments Details

C.1 2×16 Gaussians

The generative model for 2×16 Gaussians is discussed in Section 4.1. The generator, discriminator, and classification networks used for S2M sampling consist of ReLU activations and fully connected layers of input size: 2-512-512-512. We use WGAN-GP (Wei et al., 2018) as the GANs objective. We train all models using Adam optimizer (Kingma & Ba, 2015) with a learning rate of 0.0002 with $\beta_1 = 0.5, \beta_2 = 0.999$, and a batch size of 1024. The generator is trained for $4k$ iterations, and five updates of the discriminator are performed for every update of the generator. The classification networks is trained for $50k$ iterations. We do not tune the temperature of classifiers or γ_k in this experiment.

C.2 MNIST AND FMNIST

Each MNIST and FMNIST dataset consists of a training set of $60k$ images and a test set of $10k$ images. We use 10% of the training set as the validation set. To make the training set of class$_{\text{single}}$, we distribute the images belonging to the overlapping classes equally to each corresponding class$_{\text{single}}$.

The generative models for MNIST and FMNIST are discussed in Section 4.2. As similar to the original setting of GenPU, the generator consists of ReLU activations and fully connected layers of input size: 100-256-256-784. The discriminator consists of ReLU activations and fully connected layers of input size: 784-256-256. As for the GANs objective, we follow the settings introduced by the authors for baselines, and use WGAN-GP (Wei et al., 2018) for our model. We train all generative models using Adam optimizer (Kingma & Ba, 2015) with a learning rate of 0.0001, $\beta_1 = 0.5, \beta_2 = 0.999$, and a batch size of 64. The generator is trained for $200k$ iterations, and two updates of the discriminator are performed for every update of the generator.

Classification networks used for S2M sampling are obtained from multiple branches of LeNet5 (Lecun et al., 1998) architecture. We train the classifier using Adam optimizer. For MNIST 3/5 dataset, the classifier is trained for 10 epochs with a learning rate of 0.001, and the temperature of D_r is set to 2. For MNIST and FMNIST Even dataset, the classifier is trained for 50 epochs with a learning rate of of 0.0001, and the temperature of D_v is set to 4. γ_k corresponding to the intersection set are set to 0.1 for both MNIST and FMNIST.

C.3 CIFAR-10 AND CELEBA

CIFAR-10 dataset consists of a training set of $50k$ images and a test set of $10k$ images. We use 10% of the training set as the validation set. For CelebA dataset, we follow the original partition description and resize images to 64×64 for training efficiency. To make the training set of class$_{\text{single}}$, we distribute the images belonging to the overlapping classes equally to each corresponding class$_{\text{single}}$.

The generative models for CIFAR-10 and CelebA datasets are discussed in Section 4.3. We use SNGAN ResNet (Miyato et al., 2018) architecture for all models and follow the PyTorch implementation[1]. We use projection discriminator (Miyato & Koyama, 2018) for cGANs. Following (Kaneko et al., 2019), we compute the scale and bias parameters of conditional batch norm (de Vries et al., 2017) using the class specificity as weights for cGANs*, ACGAN*, and CP-GAN. For unconditional GANs, cGANs, and ACGAN (Odena et al., 2017) models, we use hinge loss (Lim & Ye, 2017) as the GAN objective and apply spectral normalization (Miyato et al., 2018) to the discriminator. For CP-GAN, we use WGAN-GP (Wei et al., 2018) as the GAN objective without spectral normalization since using SNGAN objectives degrades the performance of CPGAN as discussed in the original paper (Kaneko et al., 2019). We train all models using Adam optimizer (Kingma & Ba, 2015) with a learning rate of 0.0002, $\beta_1 = 0.5, \beta_2 = 0.999$, and a batch size of of 64. For all models, the generator is trained for $100k$ iterations, and five updates of the discriminator are performed for every update of the generator.

Classification networks used for S2M sampling are obtained from multiple branches of MobileNetV2 (Sandler et al., 2018) architecture. We first train the classifier with only $\mathcal{L}_r$ during 200 epochs for CIFAR-10 and 30 epochs for CelebA. We use SGD optimizer with a learning rate of 0.1 and cosine annealing for this training. Then, the classifier is trained with the sum of all classification losses for $3k$ iterations. Adam optimizer with the same configuration for the generator is used for the second training of classifier. We set the temperature of classifier D_r as 0.2, 0.8, and 1.6 when the size of difference index set is 0, 1, and 2, respectively. γ_k corresponding to the intersection set are set to 0.1 for CIFAR-10 and 0.5 for CelebA.

D EXPERIMENTAL RESULTS

D.1 ABLATION STUDY

To validate the effects of the temperature T and γ in S2M sampling, we perform the ablation studies in CIFAR-10 7to3 and CelebA BMS. In Table 5, we report the averaged accuracy and FID of our

[1] https://github.com/POSTECH-CVLab/PyTorch-StudioGAN

S2M sampling with different base models: unconditional GANs (GANs), cGAN, and CP-GAN. As expected, with the proper adjustment of hyperparameters, the accuracy is greatly improved without a degradation of FID. This indicates that our S2M sampling with the proper hyperparameters can sample images from an accurate data space without trading-off diversity.

Table 5: Ablation study for the hyperparameters of our S2M sampling. By adjusting the hyperparameters, we can sample more accurate images without compromising the diversity.

Method	Metric	CIFAR-10 7to3			CelebA BMS		
		GANs	cGAN	CP-GAN	GANs	cGAN	CP-GAN
Sampling w/ actual logits	Acc. (↑)	58.86±1.23	64.00±0.39	82.42±0.23	60.56±2.44	62.18±1.99	68.14±2.19
	FID (↓)	19.04±0.56	18.60±0.44	22.95±1.05	8.62±0.91	9.73±0.79	21.55±0.68
+ scale T	Acc. (↑)	63.80±1.68	70.55±0.31	84.99±0.87	68.11±3.08	69.67±1.98	71.67±2.31
	FID (↓)	19.00±0.08	18.34±0.25	23.36±1.04	8.64±0.85	9.69±0.82	21.17±0.64
+ adjust γ	Acc. (↑)	80.30±1.02	84.44±0.66	90.72±1.33	73.92±3.01	74.54±2.79	72.99±0.66
	FID (↓)	16.79±0.25	16.36±0.25	22.54±0.94	8.53±0.86	9.76±0.74	21.30±0.62

D.2 Precision and Recall on Real dataset

To give more detailed information about quality and diversity of generated samples on real dataset, we compute F-beta score (Sajjadi et al., 2018) at $\beta = 8$ as shown in Table 6. High $F_{1/8}$ (weighted precision) means high sample quality and high F_8 (weighted recall) means high sample diversity. cGAN* gets high F_8 on CelebA BMS dataset, nonetheless, the outputs of cGAN* are not well distinguishable as seen with the low accuracy reported in Table 3. Our S2M sampling with GANs and cGANs consistently draws samples with high quality and diversity which is comparable to oracle methods.

Table 6: Results of $F_{1/8}$ and F_8 on CIFAR-10 7to3 and CelebA BMS.

Model	CIFAR-10 7to3		CelebA BMS	
	$F_{1/8}$ (↑)	F_8 (↑)	$F_{1/8}$ (↑)	F_8 (↑)
cGAN (Oracle)	0.9824±0.002	0.9739±0.002	0.9808±0.003	0.9487±0.003
ACGAN (Oracle)	0.9824±0.001	0.9744±0.002	0.9853±0.003	0.9571±0.004
cGAN*	0.9728±0.003	0.9497±0.002	0.9781±0.006	0.9616±0.005
ACGAN*	0.9660±0.001	0.9406±0.002	0.9749±0.003	0.9493±0.004
CPGAN	0.9660±0.006	0.9314±0.004	0.9237±0.009	0.9151±0.016
GANs + Ours	0.9811±0.001	0.9690±0.006	0.9845±0.003	0.9606±0.001
cGAN + Ours	0.9814±0.002	0.9708±0.005	0.9808±0.006	0.9604±0.004
CPGAN + Ours	0.9687±0.004	0.9380±0.005	0.9348±0.008	0.9084±0.017

D.3 Qualitative Results

In this section, we provide the qualitative results for the experiments on CIFAR-10 7to3 and CelebA BMS.

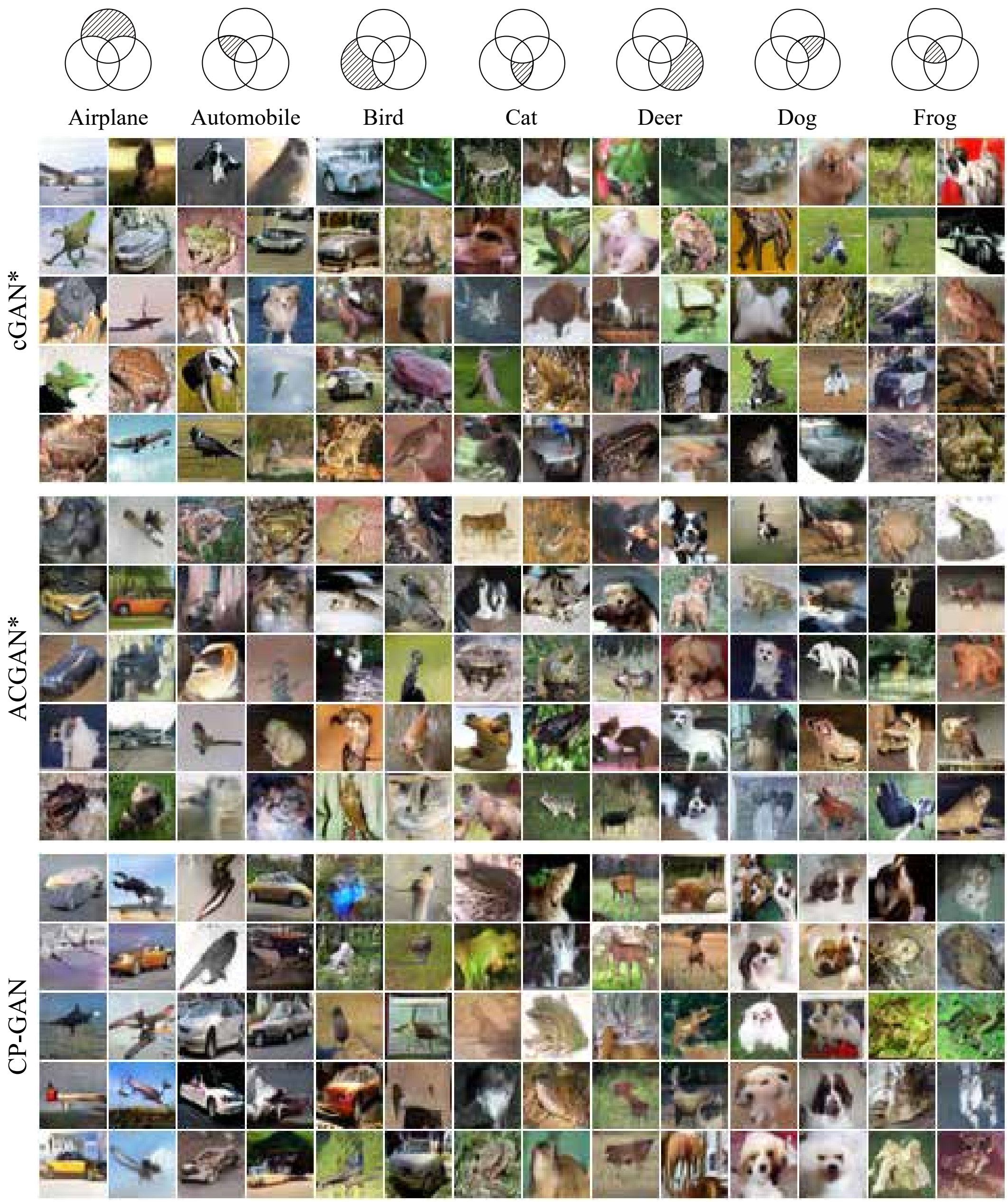


Figure 8: Qualitative results for cGAN*, ACGAN*, and CP-GAN on CIFAR-10 7to3. For cGAN* and ACGAN*, label values of $1/m$ are given for each class$_{\text{single}}$ in the intersection of m class$_{\text{single}}$.

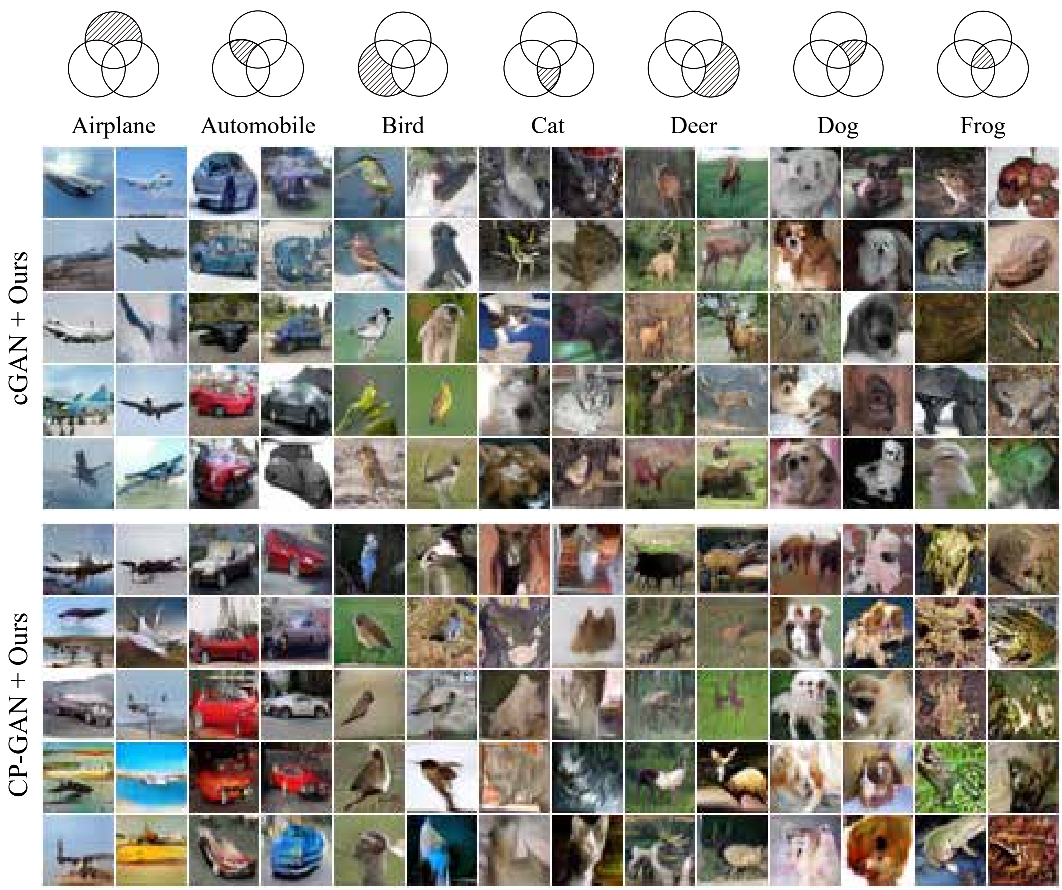


Figure 9: Qualitative results of applying our S2M sampling to cGAN and CP-GAN on CIFAR-10 7to3.

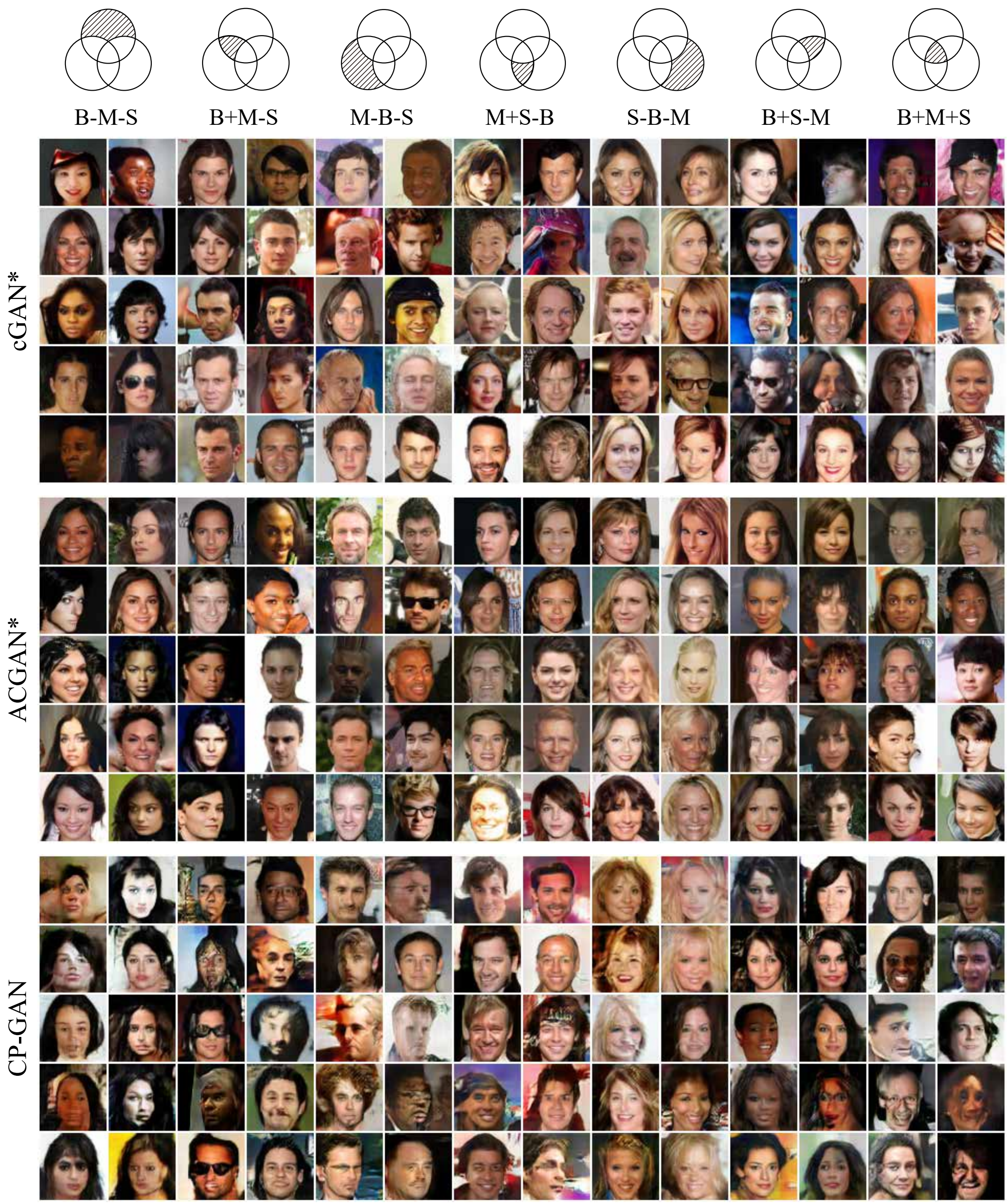


Figure 10: Qualitative results for cGAN*, ACGAN*, and CP-GAN on CelebA BMS. For cGAN* and ACGAN*, label values of $1/m$ are given for each $\text{class}_{\text{single}}$ in the intersection of m $\text{class}_{\text{single}}$. Intersections and differences are denoted by plus signs and minus signs, respectively.

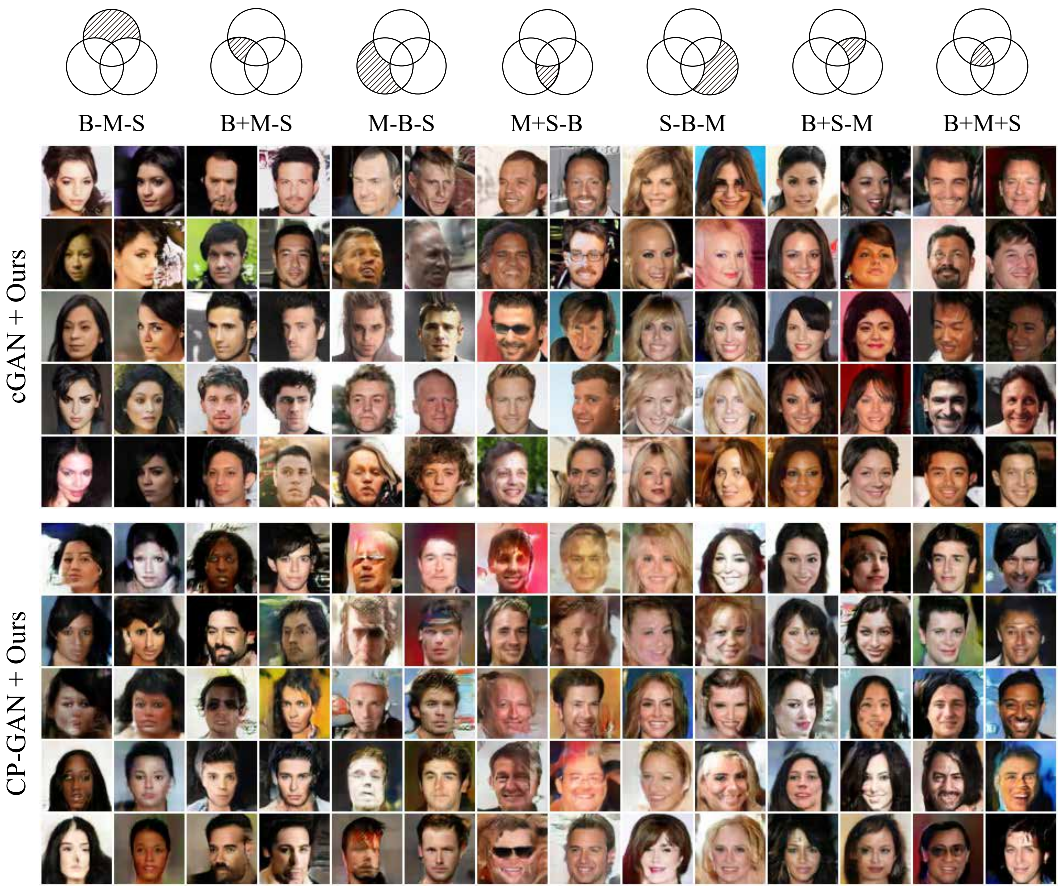


Figure 11: Qualitative results of applying our S2M sampling to cGAN and CP-GAN on CelebA BMS. Intersections and differences are denoted by plus signs and minus signs, respectively.

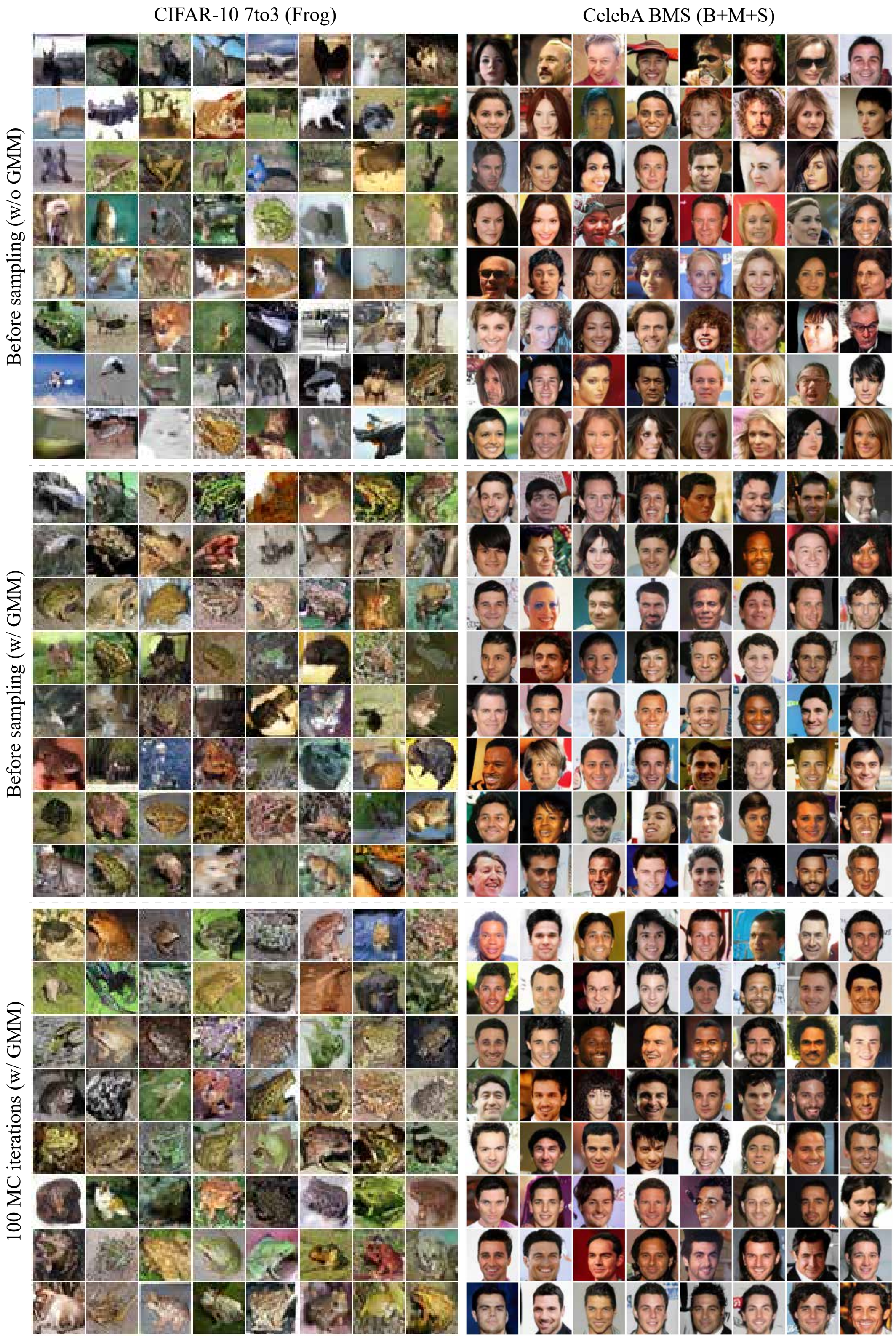


Figure 12: Qualitative results for GMM latent model discussed in Section 5.1. GMM latent model can draw samples close to target class even before sampling.