Mask-Embedded Discriminator with Region-based Semantic Regularization for Semi-Supervised Class-Conditional Image Synthesis

Yi Liu[1], Xiaoyang Huo[1], Tianyi Chen[1], Xiangping Zeng[1], Si Wu[1,2,*], Zhiwen Yu[1], and Hau-San Wong[2,*]
[1]School of Computer Science and Engineering, South China University of Technology
[2]Department of Computer Science, City University of Hong Kong
csyiliu@gmail.com, csxyhuo@mail.scut.edu.cn, csttychen@mail.scut.edu.cn, csxpzeng@gmail.com,
cswusi@scut.edu.cn, zhwyu@scut.edu.cn, cshswong@cityu.edu.hk

Abstract

Semi-supervised generative learning (SSGL) makes use of unlabeled data to achieve a trade-off between the data collection/annotation effort and generation performance, when adequate labeled data are not available. Learning precise class semantics is crucial for class-conditional image synthesis with limited supervision. Toward this end, we propose a semi-supervised Generative Adversarial Network with a Mask-Embedded Discriminator, which is referred to as MED-GAN. By incorporating a mask embedding module, the discriminator features are associated with spatial information, such that the focus of the discriminator can be limited in the specified regions when distinguishing between real and synthesized images. A generator is enforced to synthesize the instances holding more precise class semantics in order to deceive the enhanced discriminator. Also benefiting from mask embedding, region-based semantic regularization is imposed on the discriminator feature space, and the degree of separation between real and fake classes and among object categories can thus be increased. This eventually improves class-conditional distribution matching between real and synthesized data. In the experiments, the superior performance of MED-GAN demonstrates the effectiveness of mask embedding and associated regularizers in facilitating SSGL.

1. Introduction

Generative adversarial networks (GANs) [11] have made great progress in high-fidelity image synthesis [4] [14] [15] [16]. To better capture class semantics, class-conditional GANs are developed to synthesize diverse instances to match the underlying class-conditional distribution of real data [26] [30] [27] [10]. For many scenarios of real-world applications, accurate and sufficient labeled data are expensive to obtain, since the collection process typically needs expertise. The difficulty of acquiring labeled data has motivated the research on semi-supervised generative learning, which addresses a challenging task: how to train a reliable generative model for the case where there are a limited amount of labeled data together with a large amount of unlabeled data.

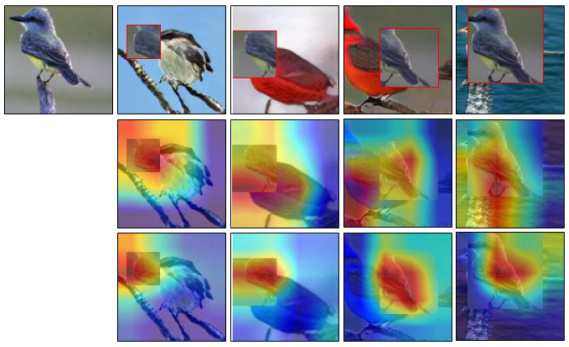


Figure 1. Visually compare the activation maps of a generic discriminator (middle row) and a mask-embedded discriminator (bottom row) used in MED-GAN on CUB-200. The top row shows a real image and the images constructed via random regional replacement between the real image and a number of fake images.

To improve class-conditional image synthesis in the semi-supervised setting, a number of techniques have been explored for enhancement on generators/discriminators. Wu et al. [43] performed class-wise mean feature matching between synthesized and real data in a classifier feature space. In [24], real and fake images of each class were randomly mixed through regional replacement. The resulting images were used to regularize a discriminator. However, a class-conditional discriminator always tends to learn the most discriminative features. As shown in Figure 1, it focuses on the regions which are not necessarily important, especially for the case of limited supervision. This may de-

*Corresponding author.

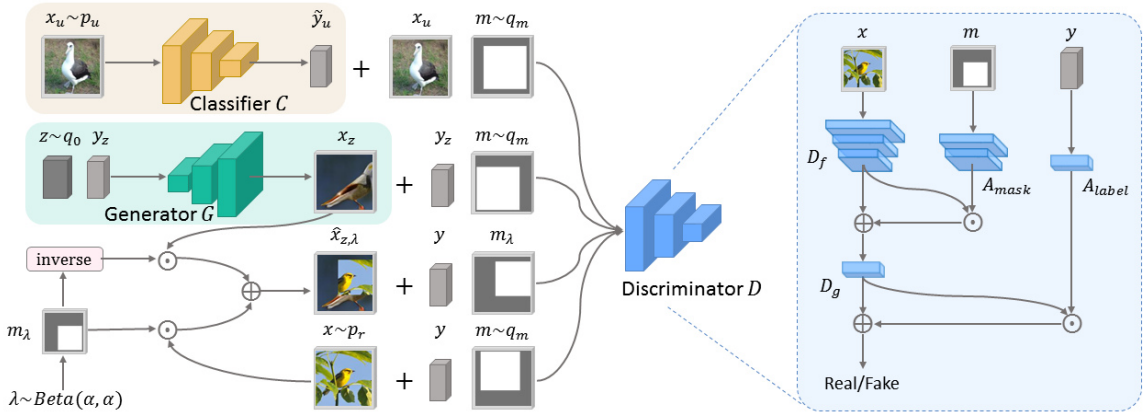


Figure 2. An overview of the proposed MED-GAN model, which consists of three constituent networks: a generator G, a discriminator D, and a classifier C. C learns to produce pseudo labels $\tilde{y}_u$ of the unlabeled data x_u, such that they can be used as well as the labeled data x. To enforce G to focus more on the synthesis of class semantics-based content, a mask embedding model A_{mask} is incorporated to induce D to discover the local differences between real and synthesized images. Also based on mask embedding, the images $\hat{x}_{z,\lambda}$ constructed via random regional replacement between x and synthesized images x_z can be used to increase the degree of class separation in the discriminator feature space, which benefits class-conditional distribution matching between real and synthesized data.

grade the generation performance of the class-conditional generator. In this work, we will explore how to induce a discriminator to distinguish between real and synthesized data from a local point of view. By reducing the dependence of the discriminator on the background, the generator may focus more on the synthesis of class semantics-related content.

More specifically, we propose a semi-supervised Generative Adversarial Network with a Mask-Embedded Discriminator (MED-GAN) for facilitating class-conditional image synthesis. Figure 2 illustrates the structure of MED-GAN. Considering the effectiveness of random regional replacement in constructing complex images, we adopt this strategy to combine real and synthesized data. As for training the discriminator, there is a lack of guideline on how to assign source (real/fake) labels to the resulting instances. To bypass this issue, we incorporate a mask embedding module in the discriminator, such that the mask of regional replacement can be embedded into the discriminator feature space. This module aims to associate spatial information with the corresponding features, and the discriminator thus applies more attention on the specified region. Further, we adopt a region-based consistent and contrastive regularization approach to regularize the discriminator feature space. By imposing separation between real and fake classes and among object categories, the generator needs to improve synthesis quality, while at the same time learn more precise class semantics in order to deceive the enhanced discriminator. Extensive validation experiments verify that the adopted techniques are effective for significantly improving synthesis quality in terms of Fréchet Inception Distance (FID) [13]. Moreover, MED-GAN is able to consistently outperform the previous state-of-the-art semi-supervised GANs on multiple standard benchmarks.

Main contributions. We summarize the main contributions of this work as follows: (1) To induce a generator to focus more on synthesizing class semantics-related content, we first enhance a discriminator by incorporating a mask embedding module, which learns to associate spatial information with the corresponding discriminator features. (2) Also based on mask embedding, we construct region-based regularizers to impose class separation on the discriminator feature space, which facilitates class-conditional distribution matching. (3) We judiciously design the optimization formulation of the constituent networks. A classifier is jointly trained with the generator and discriminator, such that they can mutually reinforce each other for capturing more precise class semantics. (4) In addition to semi-supervised generative learning, we apply the developed techniques to the fully-supervised case where the generation performance of a BigGAN can also be significantly improved, which indicates the applicability of our techniques to generic class-conditional GANs.

2. Related Work

2.1. Semi-Supervised Image Classification

There are a variety of techniques developed for image classification in the semi-supervised setting. To encourage high-confidence predictions, a widely used strategy is to minimize the entropy of the posterior class probability distribution on unlabeled data [12]. To smooth the decision boundaries, this strategy is often combined with other consistency-based regularization approaches. In [28], Miyato et al. proposed a virtual adversarial training method to compute an adversarial perturbation which maximal-

ly changes the model's predictions, while the model was trained to resist the perturbation at the same time. In addition, Park et al. [31] proposed a virtual adversarial dropout method to perturb the model's parameters. Laine and Aila [21] proposed a Π-model, in which different stochastic transformations were applied to each instance, while requiring the corresponding predictions to be consistent. To exploit the similarities among unlabeled instances, Luo et al. [25] built a prediction-based graph, and regularized a model by minimizing the representation divergence of the neighbors in the graph. Incorporating an additional model is also an effective way to construct consistency-based regularization [32] [44] [17]. In [37], a 'Mean Teacher' model was proposed to jointly train a student network and a teacher network. These two networks were required to have consistent predictions on unlabeled instances. Also along this line, a correction module was incorporated to enhance the complementarity between constituent networks [44]. In [2], Athiwaratkun et al. proposed a stochastic weight averaging strategy to improve the generalization performance of the consistency-based methods. More recently, Berthelot et al. [3] adopted a MixUp method [46] [38] to linearly combine training images and corresponding class labels, and enforced a model to have linear predictions along the interpolation path. Verma et al. [39] combined the 'Mean Teacher' model with MixUp, such that the teacher network was able to provide more reliable pseudo labels when mixing unlabeled data.

2.2. Semi-Supervised Image Synthesis

Semi-supervised generative learning has exhibited the possibility of class-conditional high-fidelity image synthesis, conditioned on limited supervision [19] [6] [24]. A number of GAN-based methods have been developed for this task. Due to the lack of labeled data, a widely used strategy is to train a categorical discriminative network, which aims at distinguishing real data from fake data along with predicting the class labels of real data. Toward this end, Springenberg presented a categorical generative adversarial network (CatGAN) in [35]. The discriminator in CatGAN was required to produce high-confidence predictions of the class labels of real instances, while the predicted class probability distributions of the synthesized instances should be close to uniform. In [41], Wei et al. adopted the Wasserstein GAN (WGAN) [1] to stabilize the GAN's training process. In [33], a variety of techniques were also explored to improve model stability and synthesis quality. Another effective strategy is to incorporate a classifier into the minimax game. Li et al. [22] proposed a Triple-GAN model, in which a classifier competed with a discriminator by estimating the class labels of unlabeled data as accurately as possible. To improve the class separability of synthesized data, Wu et al. [43] incorporated class-conditional distribution matching in Triple-GAN. Furthermore, Gan et al. [9] proposed a Triangle-GAN model, in which one more discriminator was used to identify two types of fake instance-label pairs: synthesized instances with specified labels and unlabeled instances with pseudo labels. To address the issue of imbalance between real and synthesized data, Liu et al. [24] applied random regional replacement [45] to construct between-class instances, and regularize the behaviors of the classifier and discriminator in Triangle-GAN. To assist image classification, in [5], a 'bad' generator was trained to synthesize the instances, which were located in the low-density regions in the classifier feature space, such that the synthesized data were complementary to the real data to a certain extent. To increase the class margin of synthesized data, Dong and Lin [7] proposed a MarginGAN model based on Triple-GAN. Unlike Triple-GAN, the generator competed with the classifier which was trained to minimize the class margin of the synthesized data.

Differences from the existing works. Although the random regional replacement method [45] was applied to image synthesis [34] [24], we adopt a fundamentally different approach to utilize the constructed data. (1) We incorporate a mask embedding module to enable a discriminator to discover local differences between real and synthesized images, which benefits the synthesis of class semantics-related content. (2) We further exploit mask embedding to construct effective regularizers to increase the degree of class separation in the discriminator feature space, which benefits class-conditional distribution matching. These two aspects distinguish MED-GAN from the existing semi-supervised GANs.

3. Preliminaries

Before describing MED-GAN in detail, we briefly introduce an image mixing method based on random regional replacement denoted by $\mathcal{T}$. Let x_i and x_j denote paired images. To mix x_i and x_j, a binary mask m_λ of spatial resolution $\mathtt{H}\times\mathtt{W}$ is constructed by determining a rectangular region and setting the values of the elements in the region to 1, and those outside the region to 0, where the random variable $\lambda \sim \mathtt{Beta}(\alpha,\alpha)$ controls the location and size of the region. Specifically, the top-left corner (u_0, v_0) of the region is randomly sampled on the image plane, and the bottom-right corner is positioned at $(u_0+\sqrt{1-\lambda}\mathtt{W}, v_0+\sqrt{1-\lambda}\mathtt{H})$. Next, the constructed image $\hat{x}_\lambda$ is represented as follows:

$$\begin{aligned}\hat{x}_\lambda &= \mathcal{T}(x_i, x_j, m_\lambda),\\ &= m_\lambda \odot x_i + (1-m_\lambda)\odot x_j,\end{aligned} \tag{1}$$

where $\odot$ denotes element-wise multiplication. $\hat{x}_\lambda$ is more complex than x_i and x_j, and inferring its class label is thus challenging, especially in the case where the original images are from different classes. Regularizing the model's

behavior on the constructed data is typically beneficial for the generalization performance.

4. Proposed Approach

4.1. Overview

In the semi-supervised setting, there are a limited amount of instance-label pairs $(x, y) \sim p_r$ and a large amount of unlabeled instances $x_u \sim p_u$, where p_r and p_u denote the distributions of labeled and unlabeled data, respectively. The proposed framework aims to train a class-conditional generative model with the limited supervision, and the synthesized instances should be indistinguishable from real instances, while at the same time hold correct class semantics. Toward this end, our framework consists of three components: a generator G, a discriminator D, and a classifier C. For class-conditional synthesis, G maps a random vector z together with a specified class label y_z to an instance $x_z = G(z, y_z)$, where z is sampled from a prior distribution q_0. On the other hand, D is trained to distinguish real pairs (x, y) from the synthesized ones (x_z, y_z). To exploit unlabeled data x_u, C learns to produce pseudo labels $\tilde{y}_u$ as accurately as possible, and the resulting pairs $(x_u, \tilde{y}_u)$ can be utilized as well as the labeled data.

To induce G to learn precise semantics of different object categories, D is enhanced by incorporating a mask embedding module A_{mask}, such that it is able to apply more attention on the specified region. An important benefit is to fully exploit the constructed data via random regional replacement. We further regularize the discriminator feature space by incorporating region-based consistent and contrastive regularizers, which benefit class-conditional distribution matching between real and synthesized data. G, D, and C are jointly optimized in the proposed framework. With the enhancement on D, G and C mutually reinforce each other during training. G learns class-conditional distributions with the help of the unlabeled data and pseudo labels estimated by C, and C can in turn learn from G by using the synthesized instance-label pairs.

4.2. Enhancement on the Discriminator

4.2.1 Mask embedding

Let m denote a binary mask, in which the values of the elements in a rectangular region are 1, and those outside the region are 0. We can use m to encode the region that needs attention. To encourage the discriminator to apply more attention on the specified region, we modify the discriminator D by including an additional mask embedding module A_{mask} on top of the backbone denoted by D_f, as depicted in Figure 2. A_{mask} is expected to offer insights on the important spatial regions when training D. By inputting m, A_{mask} embeds it into an intermediate feature space, and the resulting weight maps are represented by $A_{mask}(m)$. We further use $A_{mask}(m)$ to weight the intermediate feature maps as follows:

$$\widehat{D}_f(x, m) = D_f(x) + A_{mask}(m) \odot D_f(x), \qquad (2)$$

where $\widehat{D}_f(x, m)$ represents the mask-embedded features. Different from the attention-based methods [42] [47], A_{mask} aims to associate the discriminator features with the specified region. For identifying real and synthesized data of each class, we need another embedding module A_{label} to further embed class label at the last latent layer D_g as follows:

$$\widehat{D}_g(x, m, y) = D_g(x) + A_{label}(y) \odot D_g(x). \qquad (3)$$

The resulting features $\widehat{D}_g(x, m, y)$ are used to make a final decision.

Discussion. Mask embedding unlocks more effective approaches of utilizing the constructed images via random regional operation to regularize the discriminator. We can mix real and synthesized images as $\hat{x}_{z,\lambda} = \mathcal{T}(x, x_z, m_\lambda)$. For training D, it is crucial to determine the source (real/fake) label and class label for $\hat{x}_{z,\lambda}$, since $\hat{x}_{z,\lambda}$ is composed of a real image and a synthesized image. Constructing a soft source label via linear combination conflicts with the discriminator's role in the adversarial training process. In [24], x and x_z are required to have the same class label, and the source label of $\hat{x}_{z,\lambda}$ is simply determined according to the ratio of the real image part. However, the case of the ratio≈0.5 still confuses the discriminator. Empowered by mask embedding, we can bypass this issue, since the mixing mask m_λ can be utilized by our D. Both the source and class labels of $\hat{x}_{z,\lambda}$ are consistent with that of x, and the training objective is to maximize $D(\hat{x}_{z,\lambda}, m_\lambda, y)$, which denotes the probability of $\hat{x}_{z,\lambda}$ being from real data, conditioned on the mask and class label. By competing with the enhanced discriminator, the generator is encouraged to focus more on class semantics-related content.

4.2.2 Region-based regularization

To facilitate class-conditional distribution matching between real and synthesized data, we explore the benefits of mask embedding in regularizing the discriminator feature space. To enable D to capture class semantics-related information, the discriminator features should be robust to changes in the masks, and a consistency loss can thus be defined as follows:

$$L_{cons} = \mathbb{E}_{\substack{\hat{x}_{z,\lambda}, \hat{x}_{z,\lambda'} \sim \hat{p}_z \\ \|m_\lambda - m_{\lambda'}\| < \epsilon}} [\|\widehat{D}_f(\hat{x}_{z,\lambda}, m_\lambda) - \widehat{D}_f(\hat{x}_{z,\lambda'}, m_{\lambda'})\|_2^2], \qquad (4)$$

where $\hat{p}_z$ denotes the distribution of the constructed instances, and ϵ denotes a hyper-parameter controlling the degree of difference between the masks. Since both the real

image parts of $\hat{x}_{z,\lambda}$ and $\hat{x}_{z,\lambda'}$ are from x, minimizing L_{cons} encourages D to apply more attention to the content shared by the constructed images. This strategy exploits the large amount of synthesized data for learning class semantics.

When constructing $\hat{x}_{z,\lambda}$, we can also define a contrastive prediction task on the following pairs: $(x, \hat{x}_{z,\lambda})$ is positive since they have the same image part indicated by the mask m_λ, and (x, x_z) is negative. Our objective is to push x and $\hat{x}_{z,\lambda}$ closer together, while moving away from x_z in the feature space, and the corresponding contrastive loss function is defined as follows:

$$L_{ctrs} = \mathbb{E}_{\hat{x}_{z,\lambda}\sim\hat{p}_z}[\max(\|\widehat{D}_f(x, m_\lambda) - \widehat{D}_f(\hat{x}_{z,\lambda}, m_\lambda)\|_2^2 - \|\widehat{D}_f(x, m_\lambda) - \widehat{D}_f(x_z, m_\lambda)\|_2^2 + \gamma, 0)], \tag{5}$$

where γ denotes a margin that separates the negative instances from the positive ones. Minimizing L_{ctrs} not only enforces A_{mask} to accurately discover the features associated with the region indicated by the mask, but also induces D to learn more discriminative representation to separate real and synthesized data from a local point of view. It is noted that the class label of x is not necessarily the same as that of x_z. When $y \neq y_z$, the class separability in the feature space can also be improved by minimizing this contrastive loss. In this case, the generator is able to more effectively learn class-conditional data distributions.

4.3. Joint Training of Constituent Networks

Optimizing the generator. To increase the proximity of the synthesized instances to the real data of each class, one of G's objectives is to fool D by minimizing an adversarial training loss defined as follows:

$$L_{advs} = \mathbb{E}_{\substack{z\sim q_0 \\ m\sim q_m}}[\log(1 - D(G(z, y_z), m, y_z))], \tag{6}$$

where the mask m is sampled from a prior distribution q_m to encode a random rectangular region, e.g., various sized regions whose size is distributed evenly between 75% and 100% of the whole image area and whose aspect ratio is chosen randomly between 3/4 and 4/3. In addition, G should be penalized for synthesizing instances which deviate from the specified class. The classifier C serves to complement D on this point. We define another objective of G to synthesize instances which can be correctly recognized by C, and the corresponding recognition loss is defined as follows:

$$L_{recg} = \mathbb{E}_{z\sim q_0}[-y_z \log C(G(z, y_z))], \tag{7}$$

where $C(\cdot)$ represents the estimated class probability distribution of an instance. By minimizing L_{recg}, G is encouraged to enrich the class semantics of synthesized instances under the guidance of C. After integrating the above two aspects, the optimization problem of G is formulated as follows:

$$\min_{\theta_G} L_{advs} + L_{recg}, \tag{8}$$

where G is parameterized by θ_G. By competing with D and working cooperatively with C, G is trained to synthesize high-fidelity images, while at the same time hold identifiable class semantics.

Optimizing the classifier. To exploit real unlabeled data, C is used to estimate their class labels as accurately as possible. Due to lack of real labeled data, the synthesized instance-label pairs (x_z, y_z) are used to extend the training set. Instead of directly feeding the training instances to C, we construct two types of 'multi-label' instances via random regional replacement: $\hat{x}_{z,\lambda} = \mathcal{T}(x, x_z, m_\lambda)$ mixing between real labeled and synthesized instances, and $\hat{x}_{u,\lambda} = \mathcal{T}(x, x_u, m_\lambda)$ mixing between real labeled and unlabeled instances. Since $\hat{x}_{z,\lambda}/\hat{x}_{u,\lambda}$ is derived from two different images, its class labels can be defined as $\hat{y}_{z,\lambda} = \{y, y_z|\lambda\}/\hat{y}_{u,\lambda} = \{y, \tilde{y}_u|\lambda\}$, where λ denotes the ratio between the two labels. C is required to identify the associated classes of $\hat{x}_{z,\lambda}/\hat{x}_{u,\lambda}$, and the prediction is evaluated as follows:

$$\varphi(\hat{y}_{z,\lambda}, C(\hat{x}_{z,\lambda})) = -(1-\lambda)y \log C(\hat{x}_{z,\lambda}) - \lambda y_z \log C(\hat{x}_{z,\lambda}), \tag{9}$$

and $\varphi(\hat{y}_{u,\lambda}, C(\hat{x}_{u,\lambda}))$ can be computed in a similar way. Consequently, C can be trained in a supervised manner, and the optimization formulation is expressed as follows:

$$\min_{\theta_C} \mathbb{E}_{\hat{x}_{z,\lambda}\sim\hat{p}_z}[\varphi(\hat{y}_{z,\lambda}, C(\hat{x}_{z,\lambda}))] + \mathbb{E}_{\hat{x}_{u,\lambda}\sim\hat{p}_u}[\varphi(\hat{y}_{u,\lambda}, C(\hat{x}_{u,\lambda}))], \tag{10}$$

where C is parameterized by θ_C, and $\hat{p}_u$ denotes the distribution of the constructed instances between real labeled and unlabeled instances. Both the increased diversity and complexity of training data benefit C's generalization performance.

Optimizing the discriminator. To enforce G to precisely match the underlying class-conditional distributions, D is trained in opposition to G. There are four types of training data fed to D: real labeled data (x, y), real unlabeled data $(x_u, \tilde{y}_u)$, synthesized data (x_z, y_z), and constructed data $(\hat{x}_{z,\lambda}, y)$. Compared to x, it is more difficult for D to identify $\hat{x}_{z,\lambda}$ as real, since only a part of $\hat{x}_{z,\lambda}$ is from x. Indicated by m_λ, D can be enhanced by minimizing an identification loss on the constructed data as follows:

$$L_{idnt} = \mathbb{E}_{\hat{x}_{z,\lambda}\sim\hat{p}_z}[\log(1 - D(\hat{x}_{z,\lambda}, m_\lambda, y))]. \tag{11}$$

For (x, y) and (x_z, y_z), D is encouraged to apply more attention on discovering their differences in random rectangular regions. We formulate the optimization problem of D

as follows:

$$\max_{\theta_D} \mathbb{E}_{\substack{x\sim p_r \\ m\sim q_m}} [\log D(x, m, y)] + \mathbb{E}_{\substack{x_u\sim p_u \\ m\sim q_m}} [\log D(x_u, m, \tilde{y}_u)] + L_{advs} - L_{cons} - L_{ctrs} - \mu L_{idnt}, \tag{12}$$

where μ is a weighting factor. By providing pseudo labels for real unlabeled data, C cooperates with D in matching the synthesized data with real unlabeled data.

5. Experiments

We evaluate MED-GAN on multiple standard benchmarks. Our experiments mainly include four aspects: (1) We illustrate the effectiveness of the contrastive regularization in the discriminator feature space on synthetic data. (2) We also investigate the relative contributions of mask embedding and adopted regularizers on natural image synthesis. (3) We further perform extensive comparison with state-of-the-art semi-supervised generative methods in both image synthesis and classification. (4) We finally explore the applicability of the proposed improvement techniques to fully-supervised generative learning.

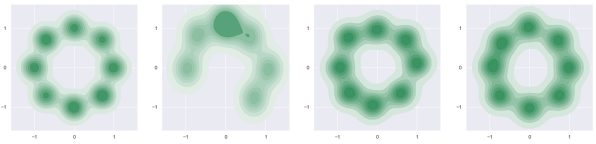
5.1. Datasets and Settings

We conduct extensive experiments on diverse datasets: CIFAR-10 and CIFAR-100 [20], FaceScrub [29], and CUB-200 [40]. CIFAR-10 (CIFAR-100) contains 50k and 10k natural images of resolution 32×32 from 10 (100) object categories for training and testing, respectively. FaceScrub is a human face dataset, in which the 100 largest classes are selected to build the FS-100 dataset. FS-100 contains about 13k training images and 2k test images of size 64×64. CUB-200 contains about 6k training images and 6k test images from 200 bird classes. In the experiments, the CUB images are resized to 128×128.

To comply with the semi-supervised setting in the literature, there are 4k, 10k, 2k, and 2.8k labeled images, and the remaining images are unlabeled in CIFAR-10, CIFAR-100, FS-100, and CUB-200, respectively. The networks are jointly trained from scratch. There are a total of 600 training epochs, and each batch contains 50/50/50 labeled/unlabeled/synthesized instances (16/16/16 for FS-100, 32/32/32 for CUB-200). The Adam optimization method [18] is adopted for stochastic gradient descent. The learning rate ς and two momentum parameters (β_1 & β_2) are set to 0.0002, 0, and 0.999, respectively. For mixing images as in Eq.(1), the random vector λ is drawn from the distribution $\texttt{Beta}(0.2, 0.2)$, which is the same as [24]. The hyper-parameter γ in Eq.(5) and the weighting factor μ in Eq.(12) are set to 0.5 and 0.1, respectively. The impact of γ and μ will be investigated in Sec.5.5. We use the Inception Score (IS) [33] and FID [13] to quantitatively evaluate the quality of synthesized images.

Table 1. The results on synthetic data.

Method	# Modes ↑	% HQ int. ↑	Reverse KL ↓
Baseline	5.3±1.1	58.9±2.0	0.184±0.011
+ L_{recg}	7.8±0.1	87.2±1.2	0.123±0.051
+ L_{ctrs}	**8.0±0.0**	**96.7±0.1**	**0.081±0.007**



(a) (b) (c) (d)

Figure 3. Visual comparison of synthesized points on the 2D-ring dataset: (a) Ground truth, (b) Baseline, (c) Baseline+L_{recg} and (d) Baseline+L_{recg}+L_{ctrs}.

Table 2. FID scores of the baseline and variants on CIFAR-10, CIFAR-100, FS-100 and CUB-200.

Method	CIFAR-10	CIFAR-100	FS-100	CUB-200
Baseline	13.69	19.52	23.18	35.04
+ Ma. emb.	9.11	10.87	17.13	23.73
+ Re. reg.	6.87	8.84	14.95	17.94
Improvement	***-6.82***	***-10.68***	***-8.23***	***-17.10***

5.2. Verification on Synthetic Data

We verify the effectiveness of the proposed improvement strategies on a 2D-ring synthetic dataset. There are 8 2D Gaussians with means $(\cos(2k\pi/8), \sin(2k\pi/8))$, $k \in \{0, \ldots, 7\}$ and the same variance 0.1. The training set is built by sampling 16 labeled points and 256 unlabeled points from each Gaussian. Similar to MED-GAN, a baseline model consists of a generator, a discriminator, and a classier, and all of them are implemented by multi-layer perceptrons. Note that there is no mask embedding in the discriminator due to the non-image data. We construct positive and negative pairs according to (pseudo) class labels. The baseline is trained based on Triple-GAN [22]. We incorporate the recognition loss in Eq.(7) and contrastive loss in Eq.(5) to improve the baseline, and the resulting models are referred to as 'Baseline + L_{recg}' and 'Baseline + L_{recg} + L_{ctrs}'. We adopt three metrics: the number of covered modes, percentage of high quality (HQ) instances and reverse KL divergence, to evaluate the performance of the baseline and variants as [23] [36]. The results shown in Table 1 and the density plots in Figure 3 suggest that inclusion of L_{recg} and L_{ctrs} leads to capturing all 8 modes and synthesizing higher quality instances. The insights gained on the synthetic data also apply to the real data.

5.3. Effectiveness of the Improvement Strategies

To further verify the benefit of the mask embedding and related regularization components, we build a baseline

Table 3. IS and FID scores of MED-GAN and state-of-the-art semi-supervised GANs on CIFAR-10, CIFAR-100, FS-100 and CUB-200.

	CIFAR-10 (4k)		CIFAR-100 (10k)		FS-100 (2k)		CUB-200 (2.8k)	
Method	IS↑	FID↓	IS↑	FID↓	IS↑	FID↓	IS↑	FID↓
ImprovedGAN [33]	5.56±0.28	47.25	-	-	-	-	-	-
Triple-GAN [22]	5.77±0.14	47.08	-	-	-	-	3.91±0.05	140.94
Triangle-GAN [9]	6.56±0.07	35.31	-	-	-	-	4.22±0.03	96.42
ETGAN [43]	7.23±0.09	25.64	4.86±0.04	65.11	1.57±0.02	57.58	3.95±0.06	133.57
R^3-CGAN [24]	7.42±0.05	20.34	7.49±0.01	26.29	1.73±0.02	25.28	4.46±0.08	88.62
MED-GAN	**8.47±0.08**	**5.76**	**9.23±0.12**	**8.06**	**1.96±0.03**	**14.42**	**5.54±0.10**	**16.90**
Ground truth	*9.07±0.14*	-	*11.40±0.13*	-	*2.44±0.04*	-	*17.73±4.87*	-

model, in which the constituent networks have the same network architectures as those in our full model. The baseline is trained based on Triple-GAN. We incrementally incorporate the components to investigate the improvement of synthesis quality in terms of FID. The experiments are performed on all the four datasets, and the results of the baseline and its variants with different strategies are reported in Table 2. When incorporating mask embedding 'Ma. emb.' in the discriminator, we find that the synthesis quality can be significantly improved over 'Baseline'. In particular, the improvement reaches 6.05 FID points on FS-100. After incorporating the region-based regularization 'Re. reg.', we can further significantly improve the fidelity of synthesized images. The corresponding variant consistently outperforms the baseline by a large margin across the datasets. On CUB-200, the FID score of synthesized images decreases from 35.04 to 17.94. The result again verifies the benefit of regularizing the discriminator feature space in facilitating semi-supervised generative learning.

5.4. Comparison with State-of-the-arts

Image synthesis. We compare MED-GAN with state-of-the-art semi-supervised generative models, including ImprovedGAN [33], Triple-GAN [22], Triangle-GAN [9], ETGAN [43], and R^3-CGAN [24]. Table 3 summarizes the results of the competing methods. We also report the IS score of real images as an upper bound. R^3-CGAN outperforms other competing methods, while MED-GAN is able to achieve higher synthesis quality than R^3-CGAN on each dataset. On the common benchmarks CIFAR-10 and CIFAR-100, MED-GAN improves the previous state-of-the-art results from 7.42/20.34 and 7.49/26.29 to 8.47/5.76 and 9.23/8.06 in IS/FID, respectively. Furthermore, we notice that our achieved results are close to the ground truth on CIFAR-10 and FS-100. On the more challenging dataset CUB-200, MED-GAN performs much better than the competing methods. In Figure 4, we show a number of the images synthesized by R^3-CGAN and MED-GAN. We believe that the proposed mask embedding and regularization strategies are useful for inducing the discriminator to discover the subtle differences among the fine-grained classes.

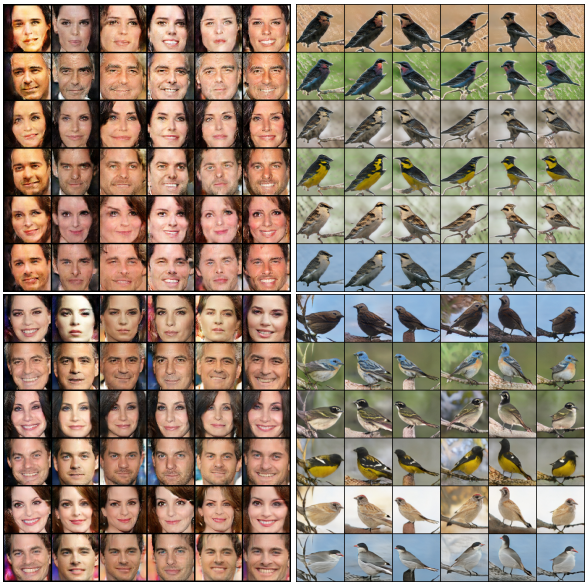


Figure 4. Examples of the images synthesized by R^3-CGAN (*upper part*) and MED-GAN (*bottom part*) on the benchmarks which from left to right are FS-100 and CUB-200. Each column shares the same random vector, and each row uses the same class label.

Table 4. Test error rates (%) of MED-GAN and state-of-the-art methods on CIFAR-10, CIFAR-100, and FS-100.

Method	CIFAR-10	CIFAR-100	FS-100
CatGAN [35]	19.58±0.58	-	-
ImprovedGAN[33]	18.63±2.32	-	-
ALI [8]	17.99±1.62	-	-
Triple-GAN [22]	16.99±0.36	-	-
Triangle-GAN [9]	16.80±0.42	-	-
GoodBadGAN [5]	14.41±0.03	-	-
CT-GAN [41]	9.98±0.21	-	-
ETGAN [43]	9.42±0.22	36.18±0.37	16.08±0.24
R^3-CGAN [24]	6.69±0.28	32.66±0.21	6.96±0.43
MarginGAN [7]	6.44±0.10	-	-
MED-GAN	**6.02±0.08**	**30.67±0.15**	**5.71±0.09**

Image classification. In MED-GAN, the classifier is jointly trained with the generator, and we also compare the resulting model with the current semi-supervised VAE/GAN-based methods in image classification. Table

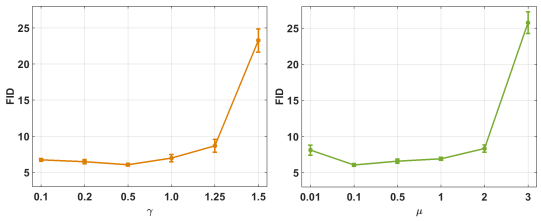


Figure 5. The impact of the margin γ (*left*) and weighting factor μ (*right*) on the synthesis quality of MED-GAN on CIFAR-10.

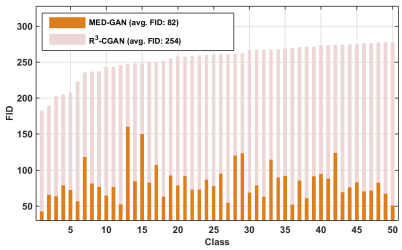


Figure 6. Comparison of MED-GAN and R^3-CGAN in terms of class-wise FID scores on CUB-200.

4 summarizes the test error rates of the competing methods on CIFAR-10, CIFAR-100 and FS-100. CIFAR-10 is a relatively simple dataset, MED-GAN, R^3-CGAN and MarginGAN achieve comparable classification results. For CIFAR-100 and FS-100, it becomes difficult to synthesize realistic images of 100 categories. The classification performance of ETGAN and R^3-CGAN are affected by the synthesis quality. On these two datasets, MED-GAN reduces the previous best test error rates of 32.66% and 6.96% (obtained by R^3-CGAN) to 30.67% and 5.71%, respectively. The results suggest that the images synthesized by MED-GAN hold more precise class semantics than those synthesized by the competing methods.

5.5. Further Analysis

Impact of parameters. We investigate the impact of the margin γ in Eq.(5) and the weighting factor μ in Eq.(12) on the synthesis quality. The experiments are conducted on CIFAR-10, and the search is limited to $\gamma = \{0.1, 0.2, 0.5, 1, 1.25, 1.5\}$ and $\mu = \{0.01, 0.1, 0.5, 1, 2, 3\}$. Figure 5 shows the changes in the FID scores of the synthesized images with different values of each parameter. In Figure (a), we observe that the synthesis quality is relatively stable when $\gamma < 1$. Figure (b) shows that the proposed approach achieves the best performance when μ is set to 0.1. Considering the complexity of the mixed instances, relatively smaller value of μ benefits the stability of the adversarial learning process.

Class-wise FID. It is important for class-conditional image synthesis to measure the extent to which synthesized instances match with the real data distribution of each class. We compare MED-GAN and R^3-CGAN in terms of class-wise FID on CUB-200. Based on the result of R^3-CGAN, we select the 50 classes with the lowest FID scores, and show the improvement achieved by the proposed approach in Figure 6. One can find that the improvement reaches 172 points in terms of average FID over the 50 classes.

Applicability. We consider that the proposed mask embedding and related regularization components can also be applied to enhance other GAN-based generative models. To verify this point, we adopt a fully-supervised BigGAN as a strong baseline, and incrementally apply the components to regularize its discriminator. Table 5 summarizes the generation performance of the resulting models in terms of FID. The result suggests that inclusion of each component can lead to consistent improvement across the four datasets accordingly, which demonstrates the possibility of its application to generic class-conditional image synthesis.

Table 5. FID scores of BigGAN and its variants on CIFAR-10, CIFAR-100, FS-100 and CUB-200.

Method	CIFAR-10	CIFAR-100	FS-100	CUB-200
BigGAN	9.06	10.32	20.76	25.62
+ Ma. emb.	7.43	8.76	18.31	21.84
+ Re. reg.	5.01	7.22	14.28	15.40
Improvement	***-4.05***	***-3.1***	***-6.48***	***-10.22***

6. Conclusion

To facilitate semi-supervised class-conditional image synthesis, our work focuses on enhancing the discriminator in a GAN-based model. We first incorporate a mask embedding module in the discriminator to associate the discriminator features with spatial information. When distinguishing real images from synthesized images, the discriminator is able to focus more on the specified regions. In this case, the generator is enforced to synthesize instances holding more precise class semantics. Under the help of mask embedding, we can more effectively exploit the constructed images via random regional replacement between real and synthesized images, and further regularize the discriminator feature space to increase the degree of class separation. The regularization of these aspects leads to significant improvement in synthesis quality.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Project No. 62072189), in part by the Research Grants Council of the Hong Kong Special Administration Region (Project No. CityU 11201220), and in part by the Natural Science Foundation of Guangdong Province (Project No. 2020A1515010484).

References

[1] Martin Arjovsky, Soumith Chintala, and Leon Bottou. Wasserstein generative adversarial networks. In *Proc. International Conference on Machine Learning*, 2017. 3

[2] Ben Athiwaratkun, Marc Finzi, Pavel Izmailov, and Andrew Gordon Wilson. There are many consistent explanations of unlabeled data: why you should average. In *Proc. International Conference on Learning Representation*, 2019. 3

[3] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. MixMatch: a holistic approach to semi-supervised learning. In *Proc. Neural Information Processing Systmes*, 2019. 3

[4] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *Proc. International Conference on Learning Representation*, 2018. 1

[5] Zihang Dai, Zhilin Yang, Fan Yang, William W. Cohen, and Ruslan Salakhutdinov. Good semi-supervised learning that requires a bad GAN. In *Proc. Neural Information Processing Systmes*, 2017. 3, 7

[6] Zhijie Deng, Hao Zhang, Xiaodan Liang, Luona Yang, Shizhen Xu, Jun Zhu, and Eric P. Xing. Strctured generative adversarial networks. In *Proc. Neural Information Processing Systmes*, 2017. 3

[7] Jinhao Dong and Tong Lin. MarginGAN: adversarial training in semi-supervised learning. In *Proc. Neural Information Processing Systmes*, 2019. 3, 7

[8] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially learned inference. In *Proc. International Conference on Learning Representation*, 2017. 7

[9] Zhe Gan, Liqun Chen, Weiyao Wang, Yunchen Pu, Yizhe Zhang, Hao Liu, Chunyuan Li, and Lawrence Carin. Triangle generative adversarial networks. In *Proc. Neural Information Processing Systmes*, 2017. 3, 7

[10] Mingming Gong, Yanwu Xu, Chunyuan Li, Kun Zhang, and Kayhan Batmanghelich. Twin auxiliary classifiers GAN. In *Proc. Neural Information Processing Systmes*, 2019. 1

[11] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proc. Neural Information Processing Systmes*, 2014. 1

[12] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Proc. Neural Information Processing Systmes*, 2004. 2

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, and Bernhard Nessler. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Proc. Neural Information Processing Systmes*, 2017. 2, 6

[14] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *Proc. International Conference on Learning Representation*, 2018. 1

[15] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1

[16] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1

[17] Zhenghan Ke, Daoye Wang, Qiong Yan, Jimmy Ren, and Rynson W.H. Lau. Dual student: breaking the limits of the teacher in semi-supervised learning. In *Proc. IEEE International Conference on Computer Vision*, 2019. 3

[18] Diederik Kingma and Jimmy Ba. Adam: a method for stochastic optimization. In *Proc. International Conference on Learning Representations*, 2015. 6

[19] Diederik P. Kingma, Shakir Mohamed, Danilo J. Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Proc. Neural Information Processing Systmes*, pages 3581 – 3589, 2017. 3

[20] Alex Krizhevsky and Geoffrey E. Hinton. Learning multiple layers of features from tiny images. In *Tech. Rep., Univ. Toronto, Toronto, ON, Canada*, 2009. 6

[21] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *Proc. International Conference on Learning Representation*, 2017. 3

[22] Chongxuan Li, Kun Xu, Jun Zhu, and Bo Zhang. Triple generative adversarial nets. In *Proc. Neural Information Processing Systmes*, 2017. 3, 6, 7

[23] Zinan Lin, Ashish Khetan, Giulia Fanti, and Sewoong Oh. PacGAN: the power of two samples in generative adversarial networks. In *Proc. Neural Information Processing Systmes*, 2018. 6

[24] Yi Liu, Guangchang Deng, Xiangping Zeng, Si Wu, Zhiwen Yu, and Hau-San Wong. Regularizing discriminative capability of CGANs for semi-supervised generative learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1, 3, 4, 6, 7

[25] Yucen Luo, Jun Zhu, Mengxi Li, Yong Ren, and Bo Zhang. Smooth neighbors on teacher graphs for semi-supervised learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3

[26] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. In *arXiv:1411.1784*, 2014. 1

[27] Takeru Miyato and Masanori Koyama. cGANs with projection discriminator. In *Proc. International Conference on Learning Representation*, 2018. 1

[28] Takeru Miyato, Shin-Ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979 – 1993, 2018. 2

[29] Hong-Wei Ng and Stefan Winkler. A data-driven approach to cleaning large face datasets. In *Proc. IEEE International Conference on Image Processing*, 2014. 6

[30] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier GANs. In *Proc. International Conference on Machine Learning*, 2017. 1

[31] Sungrae Park, Jun-Keon Park, Su-Jin Shin, and Il-Chul Moon. Adversarial dropout for supervised and semi-supervised learning. In *Proc. AAAI Conference on Artificial Intelligence*, 2018. 3

[32] Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan Yuille. Deep co-training for semi-supervised image recognition. In *Proc. European Conference on Computer Vision*, 2018. 3

[33] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Proc. Neural Information Processing Systmes*, 2016. 3, 6, 7

[34] Edgar Schonfeld, Bernt Schiele, and Anna Khoreva. A U-Net based discriminator for generative adversarial networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 3

[35] Jost Tobias Springenberg. Unsupervised and semi-supervised learning with categorical generative adversarial networks. In *Proc. International Conference on Learning Representation*, 2016. 3, 7

[36] Akash Srivastava, Lazar Valkov, Chris Russell, Michael U. Gutmann, and Charles Sutton. VEEGAN: reducing model collapse in GANs using implicit variational learning. In *Proc. Neural Information Processing Systmes*, 2017. 6

[37] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. In *Proc. Neural Information Processing Systmes*, 2017. 3

[38] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, Aaron Courville, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: better representations by interpolating hidden states. In *Proc. International Conference on Machine Learning*, 2019. 3

[39] Vikas Verma, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. In *Proc. International Joint Conference on Artificial Intelligence*, 2019. 3

[40] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 Dataset. In *Technical Report CNS-TR-2011-001, California Institute of Technology*, 2011. 6

[41] 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. International Conference on Learning Representation*, 2018. 3, 7

[42] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: convolutional block attention module. In *Proc. European Conference on Computer Vision*, 2018. 4

[43] Si Wu, Guangchang Deng, Jichang Li, Rui Li, Zhiwen Yu, and Hau-San Wong. Enhancing TripleGAN for semi-supervised conditional instance synthesis and classification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 3, 7

[44] Si Wu, Jichang Li, Cheng Liu, Zhiwen Yu, and Hau-San Wong. Mutual learning of complementary networks via residual correction for improving semi-supervised classification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3

[45] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: regularization strategy to train strong classifiers with localizable features. In *Proc. International Conference on Computer Vision*, 2019. 3

[46] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *Proc. International Conference on Learning Representation*, 2018. 3

[47] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *arXiv:1705.05512*, 2018. 4