AdaptiveMix: Improving GAN Training via Feature Space Shrinkage

Haozhe Liu[1†], Wentian Zhang[2†], Bing Li[1✉], Haoqian Wu[3], Nanjun He[2], Yawen Huang[2]
Yuexiang Li[2✉], Bernard Ghanem[1], Yefeng Zheng[2]
[1] AI Initiative, King Abdullah University of Science and Technology (KAUST),
[2]Jarvis Lab, Tencent, [3]YouTu Lab, Tencent
{haozhe.liu;bing.li;bernard.ghanem}@kaust.edu.sa; zhangwentianml@gmail.com;
{linuswu;yawenhuang;nanjunhe;vicyxli;yefengzheng}@tencent.com

Abstract

Due to the outstanding capability for data generation, Generative Adversarial Networks (GANs) have attracted considerable attention in unsupervised learning. However, training GANs is difficult, since the training distribution is dynamic for the discriminator, leading to unstable image representation. In this paper, we address the problem of training GANs from a novel perspective, i.e., *robust image classification. Motivated by studies on robust image representation, we propose a simple yet effective module, namely AdaptiveMix, for GANs, which shrinks the regions of training data in the image representation space of the discriminator. Considering it is intractable to directly bound feature space, we propose to construct hard samples and narrow down the feature distance between hard and easy samples. The hard samples are constructed by mixing a pair of training images. We evaluate the effectiveness of our AdaptiveMix with widely-used and state-of-the-art GAN architectures. The evaluation results demonstrate that our AdaptiveMix can facilitate the training of GANs and effectively improve the image quality of generated samples. We also show that our AdaptiveMix can be further applied to image classification and Out-Of-Distribution (OOD) detection tasks, by equipping it with state-of-the-art methods. Extensive experiments on seven publicly available datasets show that our method effectively boosts the performance of baselines. The code is publicly available at* https://github.com/WentianZhang-ML/AdaptiveMix.

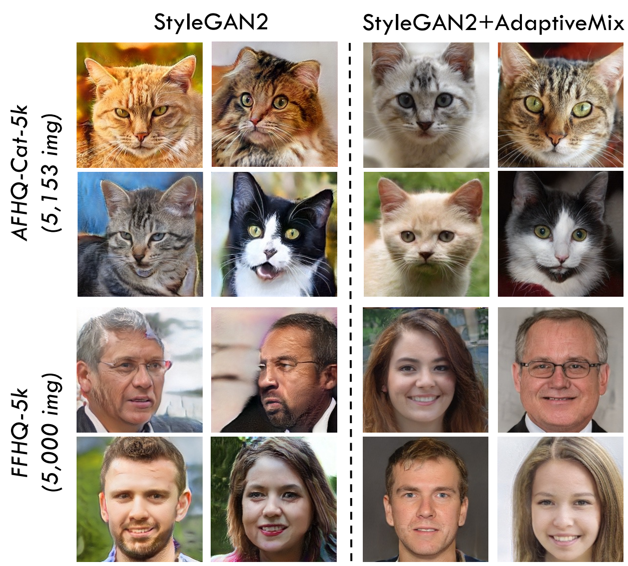


Figure 1. Results generated by StyleGAN-V2 [20] and our method (StyleGAN-V2 + AdaptiveMix) on AFHQ-Cat and FFHQ-5k. We propose a simple yet effective module AdaptiveMix, which can be used for helping the training of unsupervised GANs. When trained on a small amount of data, StyleGAN-V2 generates images with artifacts, due to unstable training. However, our AdaptiveMix effectively boosts the performance of StyleGAN-V2 in terms of image quality.

1. Introduction

Artificial Curiosity [40, 41] and Generative Adversarial Networks (GANs) have attracted extensive attention due to their remarkable performance in image generation [18,45,55,57]. A standard GAN consists of a generator and a discriminator network, where the discriminator is trained to discriminate real/generated samples, and the generator aims to generate samples that can fool the discriminator. Nevertheless, the training of GANs is difficult and unstable, leading to low-quality generated samples [23, 34].

Many efforts have been devoted to improving the training of GANs (*e.g.* [1, 5, 12, 25, 27, 34, 39]). Previous studies [37] attempted to co-design the network architecture of the generator and discriminator to balance the iterative training. Following this research line, PG-GAN [16] gradually trains

† Equal Contribution
Corresponding Authors: Bing Li and Yuexiang Li.

the GANs with progressive growing architecture according to the paradigm of curriculum learning. More recently, data augmentation-based methods, such as APA [15], ADA [17], and adding noises into the generator [20], were further proposed to stabilize the training of GANs. A few works address this problem on the discriminator side. For example, WGAN [1] proposes to enforce a Lipschitz constraint by using weight clipping. Instead, WGAN-GP [6] directly penalizes the norm of the discriminator's gradient. These methods have shown that revisions of discriminators can achieve promising performance. However, improving the training of GANs remains an unsolved and challenging problem.

In this paper, considering that the discriminator is critical to the training of GANs, we address the problem of training GANs from a novel perspective, *i.e.,* robust image classification. In particular, the discriminator can be regarded as performing a classification task that discriminates real/fake samples. Our insight is that controlling the image representation (*i.e.,* feature extractor) of the discriminator can improve the training of GANs, motivated by studies on robust image classification [28, 44]. More specifically, recent work [44] on robust image representation presents inspiring observations that training data is scattered in the learning space of vanilla classification networks; hence, the networks would improperly assign high confidences to samples that are off the underlying manifold of training data. This phenomenon also leads to the vulnerability of GANs, *i.e.,* the discriminator cannot focus on learning the distribution of real data. Therefore, we propose to shrink the regions of training data in the image representation space of the discriminator.

Different from existing works [15, 17], we explore a question for GANs: *Would the training stability of GANs be improved if we explicitly shrink the regions of training data in the image representation space supported by the discriminator?* To this end, we propose a module named AdaptiveMix to shrink the regions of training data in the latent space constructed by a feature extractor. However, it is non-trivial and challenging to directly capture the boundaries of feature space. Instead, our insight is that we can shrink the feature space by reducing the distance between hard and easy samples in the latent space, where hard samples are regarded as the samples that are difficult for classification networks to discriminate/classify. To this end, AdaptiveMix constructs hard samples by mixing a pair of training images and then narrows down the distance between mixed images and easy training samples represented by the feature extractor for feature space shrinking. We evaluate the effectiveness of our AdaptivelyMix with state-of-the-art GAN architectures, including DCGAN [37] and StyleGAN-V2 [20], which demonstrates that the proposed AdaptivelyMix facilitates the training of GANs and effectively improves the image quality of generated samples.

Besides image generation, our AdaptiveMix can be applied to image classification [9, 53] and Out-Of-Distribution (OOD) detection [11, 14, 50] tasks, by equipping it with suitable classifiers. To show the way of applying AdaptiveMix, we integrate it with the Orthogonal classifier in recent start-of-the-art work [52] in OOD. Extensive experimental results show that our AdaptiveMix is simple yet effective, which consistently boosts the performance of [52] on both robust image classification and Out-Of-Distribution tasks on multiple datasets.

In a nutshell, the contribution of this paper can be summarized as:

- We propose a novel module, namely AdaptiveMix, to improve the training of GANs. Our AdaptiveMix is simple yet effective and plug-and-play, which is helpful for GANs to generate high-quality images.
- We show that GANs can be stably and efficiently trained by shrinking regions of training data in image representation supported by the discriminator.
- We show our AdaptiveMix can be applied to not only image generation, but also OOD and robust image classification tasks. Extensive experiments show that our AdaptiveMix consistently boosts the performance of baselines for four different tasks (*e.g.,* OOD) on seven widely-used datasets.

2. Related Work

The Zoo of Interpolation. Since regularization-based methods can simultaneously improve the generalization and robustness with neglectable extra computation costs, this field attracts increasing attention from the community [7, 51, 54]. To some extent, Mixup [54] is the first study that introduces a sample interpolation strategy for the regularization of Convolution-Neural-Network(CNN)-based models. The virtual training sample, which is generated via linear interpolation with pair-wise samples, smooths the network prediction. Following this direction, many variants were proposed by changing the form of interpolation [7, 21, 29, 44, 51]. In this paper, we revisit the interpolation-based strategy and regard the mixed sample as a generated hard sample for shrinking features, which gives a new perspective for the application of mixing operations.

Regularization in GANs. Recently, various methods [1, 15, 34, 39] attempt to provide regularization on the discriminator to stabilize the training of the GANs. Previous studies designed several penalties, such as weight clipping [1] and gradient penalty [6] for the parameter of the discriminator. As this might degrade the capacity of the discriminator, spectral normalization is proposed to further stabilize the training via weight normalization. However, spectral normalization has to introduce extra structure, which might limit its application towards arbitrary network architectures. As a more flexible line, adversarial training [59], C-Reg [56], LC-Reg [43], ADA [17], and APA [15] are proposed in recent years, working similarly to the data augmentation. These methods suffer from the sensitivity to the severity of augmented data and have to use adaptive hyperparameter [15, 17]. Besides, Transform-Reg [35] designs transformation consistency regularization loss for image-to-image translation, which encourages the output image space to preserve local smoothness in the input space. Different from these approaches, we train a robust discriminator by shrinking the feature space. Without sacrificing the capability of representation, the proposed method can be elaborated into many networks and easily combined with existing regularization. Moreover, the proposed method can continually construct hard samples for training without too many hyperparameters, and thus can be used in various additional tasks, such as OOD detection [24, 26, 32, 48, 52] and image classification [4, 22, 46].

3. Method

In this paper, we investigate how to improve the training of GANs. We first propose a novel module named AdaptiveMix to shrink the regions of training data in the image representation space of the discriminator. Then, we show that our AdaptiveMix can encourage Lipschitz continuity, and thereby facilitate the performance of GANs. Finally, we equip our AdaptiveMix with an orthogonal classifier of the start-of-the-art OOD method in [52] to show how to use our module for OOD detection and image recognition tasks.

3.1. AdaptiveMix

Our goal is to improve the training of GANs by controlling the discriminator, which can be formulated from the perspective of robust image classification. Without loss of generality, let the discriminator consist of a *feature extractor* $\mathcal{F}(\cdot)$ and a classifier head $\mathcal{J}(\cdot)$, where the feature extractor is to extract feature from an image, and the classifier is to classify the extracted feature. Our insight is that we can improve the training of GANs by improving the representation of the feature extractor $\mathcal{F}$, motivated by studies on robust image representation [44, 60]. As observed in [44], vanilla classification networks scatter training data in their feature space, driving the classifier improperly to assign high confidences to samples that are off the underlying manifold of training data. Similarly, with such representation, it is difficult for the discriminator to learn the distribution of real data. Therefore, we propose to shrink the regions of training data in the image representation space supported by the feature extractor of the discriminator for improving the training of GANs.

We propose a module, termed AdaptiveMix, to shrink the regions of training data in the space represented by a feature extractor $\mathcal{F}$. However, it is intractable to directly capture the regions of training data in the feature space. Given training samples of a class c, our insight is that we can shrink its regions in the feature space by reducing the distance between hard and easy samples in the feature space, where hard samples are regarded as samples that are difficult for networks to classify. In other words, we argue that most hard samples are more peripheral than easy ones in the feature space formed by all training samples of a class, which leads the decision boundaries to enlarge the intra-class distance for covering the hard samples. Therefore, for class c, if we pull hard samples towards easy samples, the regions of training samples of class c can be shrunk in the feature space. Therefore, the proposed AdaptiveMix consists of two steps. First, AdaptiveMix generates hard samples from the training data. Second, our AdaptiveMix reduces the distance between hard and easy samples.

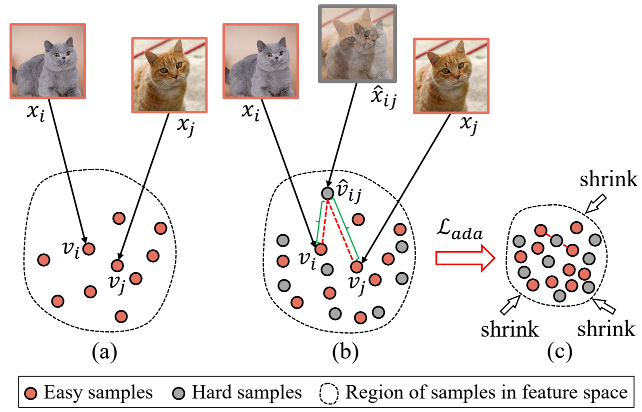


Figure 2. The illustration of our AdaptiveMix. (a) Easy samples x_i and x_j of a class are projected into the feature space by feature extractor $\mathcal{F}(\cdot)$, where $v_i = \mathcal{F}(x_i)$. (b) Hard sample $\hat{x}_{ij}$ is generated by the convex combination of a training sample pairs x_i and x_j, and is projected to the feature space. (c) AdaptiveMix shrinks the region of the class in the feature space by reducing the feature distance between easy and hard samples.

Hard Sample Generation. A naive manner of finding hard samples is to employ trained networks to evaluate training samples, where samples to which the networks assign the prediction with low confidence are considered as hard ones. However, this introduces new issues. For example, this requires well-trained networks, which are not always available. Instead, we propose a simple way to generate hard samples, inspired by the promising performance of Mixup-based image augmentation methods [7, 13, 54]. Recently, various Mixup-based methods were proposed to mix multiple images into a new image. Here we employ the vanilla version of Mixup [54] to generate hard samples for simplicity. Let $\mathcal{X} = \{x_i\}_{i=1}^N$ denote N training images, where x_i is the i-th training image. AdaptiveMix mixes a pair of training images to generate a hard sample following Mixup:

$$\hat{x}_{ij} = g(x_i, x_j, \lambda) = \lambda x_i + (1 - \lambda) x_j, \tag{1}$$

where $g(\cdot, \cdot, \lambda)$ is a function linearly combining x_i and x_j, λ is a hyper-parameter sampled from Beta distribution

$\lambda \in \mathbb{B}(\alpha, \alpha)$. As shown in Fig. 2, mixed sample $\hat{x}_{ij}$ is more confusing and difficult for networks to discriminate, compared with original training samples x_i and x_j. Without loss of generalization, we refer to an original training sample x_i as an easy sample, and $\hat{x}_{ij}$ as a hard one. Note that hard sample $\hat{x}_{ij}$ does not belong to a new class but belongs partially to the class of x_i and partially to that of x_j, according to the λ value used in Eq. 1.

AdaptiveMix Loss. Our AdaptiveMix reduces the distance between a hard sample $\hat{x}_{ij}$ and its corresponding easy ones x_i and x_j in the representation space represented by feature extractor $\mathcal{F}(\cdot)$, in order to shrink regions of training data of classes that x_i or x_j belongs to. Note that we propose a soft loss, since hard sample $\hat{x}_{ij}$ does not completely belong to the class of x_i or x_j. We reduce the distance between $\hat{x}_{ij}$ and x_i in the feature space according to the proportion of x_i in $\hat{x}_{ij}$ in the linear combination:

$$\mathcal{L}_{ada}=\sum_i \sum_j \mathbb{D}_v(\lambda \mathcal{F}(x_i)+(1-\lambda)\mathcal{F}(x_j), \mathcal{F}(\hat{x}_{ij})+\sigma), \tag{2}$$

where σ is a noise term sampled from Gaussian distribution to prevent over-fitting and $\mathbb{D}_v(\cdot,\cdot)$ refers to the metric to evaluate the distance, like L1 norm, L2 norm. Note that our AdaptiveMix does not need labels of training images; however, it is able to shrink the regions of training data for each class in the feature space (see Fig. 2), since easy sample x_i and its associated hard one $\hat{x}_{ij}$ belong to the same class.

3.2. Connections to Lipschitz Continuity

To further investigate the superiority of our method, we theoretically analyze the relationship between the proposed AdaptiveMix and Lipschitz continuity.

Preliminary. In the proposed method, the feature extractor $\mathcal{F}(\cdot)$ connects the input space $\mathcal{X}$ and the embedding space $\mathcal{V}$. Given two evaluation metrics $\mathbb{D}_x(\cdot,\cdot)$ and $\mathbb{D}_v(\cdot,\cdot)$ defined on $\mathcal{X}$ and $\mathcal{V}$, respectively, $\mathcal{F}(\cdot)$ fulfills Lipschitz continuity if a real constant K exists to ensure all $x_i, x_j \in \mathcal{X}$ meet the following condition:

$$K\mathbb{D}_x(x_i, x_j) \geq \mathbb{D}_v(\mathcal{F}(x_i), \mathcal{F}(x_j)). \tag{3}$$

Proposition. Based on the analysis in [3, 44], a flat embedding space, especially with Lipschitz continuity, is an ideal solution against unstable training and adversarial attack. Hence, the effectiveness of the proposed method can be justified by proving the equivalence between AdaptiveMix and K-Lipschitz continuity.

Theorem. Towards any K of Lipschitz continuity, AdaptiveMix is an approximate solution under L1 norm metric space.

Proof. Given x_i and x_j sampled from $\mathcal{X}$, their convex combination based pivot $\hat{x}_{ij}$ can be obtained via $g(x_i, x_j, \lambda)$. Since $\hat{x}_{ij}$ can be regarded as a sample in $\mathcal{X}$, we can transform Lipschitz continuity (Eq. (3)) to

$$\begin{aligned} K\mathcal{B} =& K(\mathbb{D}_x(\hat{x}_{ij}, \lambda x_i) + \mathbb{D}_x(\hat{x}_{ij}, (1-\lambda)x_j)) \\ &\geq \mathbb{D}_v(\hat{v}_{ij}, \lambda v_i) + \mathbb{D}_v(\hat{v}_{ij}, (1-\lambda)\lambda v_j) \\ &\geq \mathbb{D}_v(\hat{v}_{ij}, g(v_i, v_j, \lambda)) \end{aligned} \tag{4}$$

where $\mathbb{E}[\lambda] = 0.5$ and $\mathbb{E}[x_i] = \mathbb{E}[x_j]$. Hence the upper bound $\mathcal{B}$ can be estimated as 0 through mini-batch training. As $\mathbb{D}_v(\cdot,\cdot)$ is an L1-norm distance, $\mathbb{D}_v(\cdot,\cdot)$ should be no less than 0. Hence, we can get the lower and upper bound of $\mathbb{D}_v(\hat{v}_{ij}, g(v_i, v_j, \lambda))$ within Lipschitz continuity:

$$\mathbb{E}[0] \leq \mathbb{E}[\mathbb{D}_v(\hat{v}_{ij}, g(v_i, v_j, \lambda)))] \leq \mathbb{E}[K\mathcal{B}] = 0. \tag{5}$$

Therefore, if $\mathcal{F}$ is under the Lipschitz continuity, $\mathbb{D}_v(\hat{v}_{ij}, g(v_i, v_j, \lambda)))$ should be zero, and the optimal result of AdaptiveMix is identical to $\mathbb{D}_v(\hat{v}_{ij}, \Lambda_\lambda(v_i, v_j))$. Therefore, K-Lipschitz continuity can be ensured by minimizing AdaptiveMix.

Intuition. This theoretical result is also consistent with intuition. Our general idea is to shrink the feature space for robust representation. As Lipschitz continuity requires that the distance in embedding space should be lower than that in image space, shrinking feature space should be a reasonable way to approximately ensure Lipschitz continuity.

3.3. AdaptiveMix-based Image Generation

Based on the previous analysis, the proposed module can help to stabilize the training of GANs. In this paper, we show how to apply AdapativeMix to image generation by integrating it with two state-of-the-art image generation methods, WGAN [1], and StyleGAN-V2 [20]. We mainly elaborate on the integration of WGAN in the main paper, and that of StyleGAN-V2 is given in the *supplementary materials*. Thanks to the plug-and-play property of AdaptiveMix, we equip WGAN with AdaptiveMix in a simple manner. In particular, we apply AdaptiveMix to WGAN's discriminator consisting of feature extractor $\mathcal{F}(\cdot)$ and classifier head $\mathcal{J}(\cdot)$ We then rewrite the learning objective to add the AdaptiveMix to WGANs:

$$\begin{aligned} \min_G \max_{\mathcal{F},\mathcal{J}} \mathbb{E}_{x\sim p_r}[\mathcal{J}(\mathcal{F}(x))] - \mathbb{E}_{z\sim p_z}[\mathcal{J}(\mathcal{F}(G(z)))] \\ + \min_{\mathcal{F}} \mathbb{E}_{x\sim p_r, p_g}[\mathcal{L}_{ada}] \end{aligned} \tag{6}$$

where z is the noise input of the generator $G(\cdot)$; L2 norm is adapted as the metrics for $\mathcal{L}_{ada}$, the output of $\mathcal{J}(\cdot)$ refers to a scalar to estimate the realness of the given sample. To simplify the structure of $\mathcal{J}(\cdot)$, we directly adopt averaging operator as $\mathcal{J}(\cdot)$. In this paper, AdaptiveMix generates hard samples by the convex combination of real samples and fake ones generated by the generator. Such mixing is a kind of feature smoothing that enforces the decision boundaries of the discriminator to be smooth, improving the training stability of GANs. The pseudo-code is given in *supplementary material*. Note that the traditional mixing-based methods do **not** works for the zoo of WGANs, since WGAN plays a dynamic min-max game where the output of the discriminator ranges from $(-\infty, +\infty)$, while our method improves the training of WGAN.

3.4. AdaptiveMix-based Image Classification

Besides image generation, the proposed AdaptiveMix can be applied to image classification. Here we show how to apply AdativeMix to this task. Different from the image generation task, image classification requires features extracted by the feature extractor $\mathcal{F}(\cdot)$ of a classification model to be discriminative as much as possible. Our AdaptiveMix shrinks regions of training data in the feature space, which smooths features to some extent if AdaptiveMix is solely applied. Nevertheless, this can be easily addressed by adopting a proper classifier that enforces features of different classes to be separable.

Inspired by image classification method [52], we employ an orthogonal classifier $\tilde{\mathcal{J}}(\cdot)$ to ensure the class-aware separation in the feature space, where the orthogonal classifier $\tilde{\mathcal{J}}(\cdot)$ consists of several weight vectors $w_k \in \mathcal{W}$, and w_k corresponding to the k-th class. In particular, we replace the last fully-connected layer of a CNN-based classification model with the orthogonal classifier $\tilde{\mathcal{J}}(\cdot)$. Thus, given x_i, the prediction score y^k to k-th class is calculated as:

$$y_i^k = \frac{w_k^T v_i}{||w_k||\ ||v_i||}. \tag{7}$$

where $v_i = \mathcal{F}(x_i)$ denoted x_i's feature extracted by feature extractor $\mathcal{F}(\cdot)$ of the classification model. The probability p_i^k that x_i belongs to k-th class is calculated via a softmax layer:

$$p_i^k = \frac{exp(y_i^k)}{\sum\limits_{1\leq l\leq n} exp(y_i^l)} \tag{8}$$

where the set $\mathcal{P}_i = \tilde{\mathcal{J}}(v_i) = \{p_i^k | 1 \leq k \leq n\}$ forms the final output of the CNN-based model for an n-class recognition task.

By removing the bias and activation function in the last layer, the classification model maps x into the allowed norm ball space, which ensures that features corresponding to different classes can be separable. To further strengthen the class-aware separation, we then introduce the orthogonal constraint to initialize $\mathcal{W}$, which is defined as:

$$\prod_{w_k, w_l \in \mathcal{W}, k\neq l} w_k^T w_l = 0. \tag{9}$$

In addition, besides AdaptiveMix loss, we can use mixing-based cross-entropy loss in the learning objective of image classification following augmentation [54], since we use Mixup to generate hard samples (See *supplementary material* for more details on classification).

Table 1. Summary of improvements by using our AdaptiveMix, where **Gain** refers to our improvement over the baselines. Our method AdaptiveMix boosts the performance of six baselines across four tasks on seven widely-used datasets. Detailed comparison results are provided in tables specified in the *Tab. # column*.

	Dataset	Tab. #	**Gain**
Image Generation	C-10 [22]	Tab. 4	**-20.0% FID ↓**
	CelebA [30]	Tab. 4	**-54.0% FID ↓**
	FFHQ [19]	Tab. 5	**-4.9% FID ↓**
	AFHQ-CAT [2]	Tab. 5	**-43.5% FID ↓**
	FFHQ-5k [19]	Tab. 6	**-47.9% FID ↓**
Image Classification	C-10 [22]	Tab. 11	**+0.7% Acc. ↑**
	C-100 [22]	Tab. 11	**+1.5% Acc. ↑**
	T-ImageNet [4]	Tab. 11	**+5.87% Acc. ↑**
	ImageNet [38]	Tab. 11	**+1.9% Acc. ↑**
Robust Classification	C-10 [22]	Tab. 7	**+4.6× Acc. ↑**
	C-100 [22]	Tab. 8	**+5.2× Acc. ↑**
	T-ImageNet [4]	Tab. 8	**+1.1 × Acc. ↑**
OOD Detection	Benchmark [52]	Tab. 12	**+3.5% F1 ↑**

3.5. AdaptiveMix-based OOD Detection

AdaptiveMix can be easily integrated into the state-of-the-art OOD detection model of [52]. Given all training samples $x \in \mathcal{X}$ as input, we can obtain the corresponding representation $v \in \mathcal{V}$ via the trained $\mathcal{F}(\cdot)$. Then, the representative representation v_k^* of the k-th class can be obtained by computing the first singular vectors of $\{\mathcal{F}(x_i)|x_i \in \mathcal{X} \ \arg\max \widetilde{y}_i = k\}$. Note that v_k^* is calculated by SVD rather than $\mathcal{F}(x_i)$. Given a test sample x_t, the probability ϕ_t that x_t is OOD is calculated as:

$$\phi_t = \min_k \arccos(\frac{|\mathcal{F}^T(x_t)v_k^*|}{||\mathcal{F}(x_t)||}), \tag{10}$$

where x_t is categorized as an OOD sample if ϕ_t is larger than a predefined threshold ϕ^*.

4. Experiments

To evaluate the performance of our method, we conduct extensive experiments on various tasks, including image generation, image recognition, robust image classification, and OOD detection. Tab. 1 shows that the proposed method improves the baselines significantly on these tasks. Below, we first evaluate the performance of the proposed method on image generation and then test the proposed method on visual recognition tasks such as image classification and OOD detection. Note that we elaborate on datasets used in the experiments, additional experiments, and implementation details in the *supplementary material*.

Table 2. The ablation study of the proposed method on CIFAR10 [22]. Mean FID ± S.D. refers to the mean and standard deviation of FID scores when the models are trained by 100, 400, and 800 epochs. Min FID is the optimal result during training.

	Mean FID ± S.D. ↓	Min FID ↓
Baseline	106.81 ± 21.79	55.96
Baseline + AdaptiveMix	**39.52 ± 7.85**	**30.85**

Table 3. The evaluation for the Lipschitz continuity on FFHQ-5k [19] with StyleGAN-V2 [20]. The criterion refers to the averaging of $\frac{\text{embedding distance}}{\text{image distance}}$ on the given pairs of samples, *i.e.* $\frac{\mathbb{D}_v(\mathcal{F}(x_i),\mathcal{F}(x_j))}{\mathbb{D}_x(x_i,x_j)}$. Smaller is better.

	Real Samples	Generated Samples	Both
StyleGAN-V2	3.167	0.263	0.735
StyleGAN-V2 + Ours	**1.391**	**0.166**	**0.291**

4.1. Performance on Image Generation

Ablation Study. To quantify the contribution of AdaptiveMix, we test the image generation performance with or without AdaptiveMix. As listed in Tab. 2, the proposed method outperforms the baseline [1] significantly in all cases. Such improvement indicates that the proposed method boosts the training of GANs.

In addition to the quantitative analysis, we also conduct visualizations to showcase the effectiveness of the proposed method. As shown in Fig. 3, we generate two toy datasets, where each consists of 50k 2D points with the x and y coordinates of a point in the range of [-2, 2]. The distribution of points generated by our method is most similar to that of the real ones, indicating the best generation performance than the other two baselines. For further investigation, the output of the discriminator for each position is visualized in the second row of Fig. 3. If the discriminator enlarges the distance between real and generated samples like Std-GAN [5], the generator is hard to derive useful guidance from the discriminator, leading to poor generated results or mode collapse. In contrast, the proposed method shrinks the input samples; hence the confidence score map is flattened, which can provide more health gradient for the generator, resulting in a better generation performance. We also quantify such a phenomenon in the practical case. As shown in Tab. 3, we calculate the averaging ratios between the distances in the feature space and image space. By adopting AdaptiveMix, the corresponding ratios are minimized among all the cases, which can be regarded as a guarantee for the Lipschitz continuity.

Comparing with Existing Methods. To further show the superiority of the proposed method, we compare the performance of the proposed method with the well-known loss functions on the toys dataset, CIFAR10 [22], and CelebA [31]. As shown in Tab. 4, the proposed method outperforms other methods in both datasets by a large margin. Compared with the recent method, Realness GAN, a 10.81% improvement in FID is achieved by the proposed method on CIFAR10. Similarly, in the case of CelebA, the FIDs of Realness GAN and the proposed method are 23.51 and 12.43, respectively, which convincingly shows the advantage of AdaptiveMix. Note that the results in Tab. 4 are taken from [47] based on the same architecture, i.e. DCGAN. The corresponding visualization results are given in the *supplementary material*.

In order to comprehensively justify AdaptiveMix, we also compare the proposed method with the recent regularization for GANs. Tab. 5 shows the proposed method

Table 4. FIDs of DCGAN [37] using various learning objectives on CelebA [31] and CIFAR10 [22].

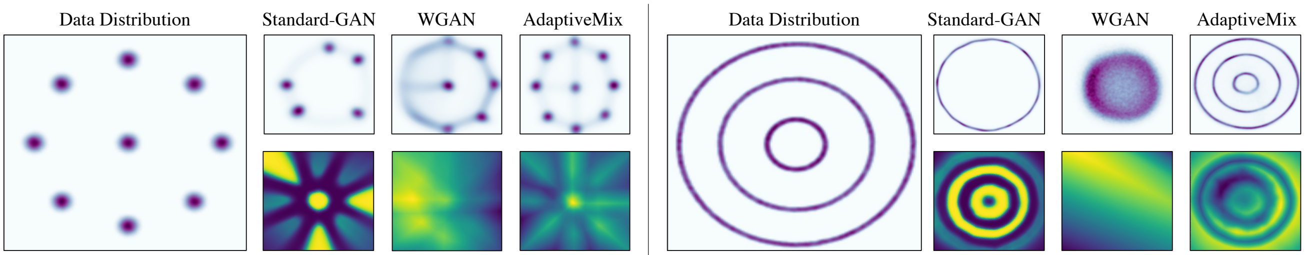
Learning Objective	CIFAR-10	CelebA
WGAN [1] (ICML'17)	55.96	-
HingeGAN [58] (ICLR'17)	42.40	25.57
LSGAN [33] (ICCV'17)	42.01	30.76
DCGAN [37] (ICLR'16)	38.56	27.02
WGAN-GP [6] (NIPS'17)	41.86	70.28
Re-implemented WGAN-GP	38.63	70.16
Realness GAN-Obj.1 [47] (ICLR'2020)	36.73	-
Realness GAN-Obj.2 [47] (ICLR'2020)	34.59	23.51
Realness GAN-Obj.3 [47] (ICLR'2020)	36.21	-
AdaptiveMix (Ours)	**30.85**	**12.43**

Table 5. FID and IS of the proposed method on AFHQ [2] and FFHQ [19] for StyleGAN-V2 [20] compared with the other state-of-the-art solutions for GAN training

Method	AFHQ-Cat-5k		FFHQ (Full)	
	FID	IS	FID	IS
StyleGAN-V2 [20] (CVPR'20)	7.737	1.825	3.862	5.243
StyleGAN-V2 (Re-Impl.)	7.924	1.890	3.810	5.185
LC-Reg [43] (CVPR'21)	6.699	1.943	3.933	**5.312**
Style GAN-V2 + Ours	**4.477**	**1.972**	**3.623**	5.222
ADA [17] (NIPS'20)	6.053	2.119	4.018	5.329
ADA (Re-Impl.)	5.582	2.059	3.713	5.200
ADA + Ours	**4.680**	**2.069**	**3.681**	**5.335**
APA [15] (NIPS'2021)	4.876	2.156	3.678	5.336
APA (Re-Impl.)	4.645	2.093	3.752	5.281
APA+Ours	**4.148**	**2.096**	**3.609**	**5.296**

Table 6. FID and IS of our method compared to previous techniques for regularizing GANs on FFHQ-5k [19]. StyleGAN-V2 [20] is used as the baseline.

Regularization	FID	IS
Baseline [20] (CVPR'20)	37.830	4.018
Baseline (Re-Impl.)	36.053	4.097
Instance Noise [42] (ICLR'17)	40.981	4.231
One-sided LS [39] (NIPS'16)	33.978	4.029
LC-Reg [43] (CVPR'21)	35.148	3.926
Ours	**18.769**	**4.332**
APA [15] (NIPS'2021)	13.249	4.487
APA (Re-Impl.)	14.368	4.855
APA+Ours	**11.498**	**4.866**



(a) Mixed Gaussian Distributions

(b) Mixed Circle Lines

Figure 3. The experimental results on a synthetic data set: 2D points from (a) nine Gaussian distributions and (b) three circles are adopted as the training data for GANs. From left to right in (a) and (b): data distribution, the samples generated by Std-GAN, WGAN, and AdaptiveMix. The first row refers to the generated results, and the second row is the corresponding confidence map of the discriminator.

can help the convergence of GANs on different datasets and achieve remarkable results. As a plug-and-play module, the proposed method can also be combined with the state-of-the-art augmentation-based methods, ADA [17] and APA [15], which further improves the generation performance of GANs (13.5% improvement in FID averagely). Finally, we evaluate the performance of the proposed method using limited training data. As shown in Tab. 6, given only 5k samples, the proposed method can significantly improve the baseline from 37.830 to 18.769 in FID. By combining with APA, AdaptiveMix can achieve the best FID (11.498) and IS (4.866) scores.

4.2. Performance on Visual Recognition

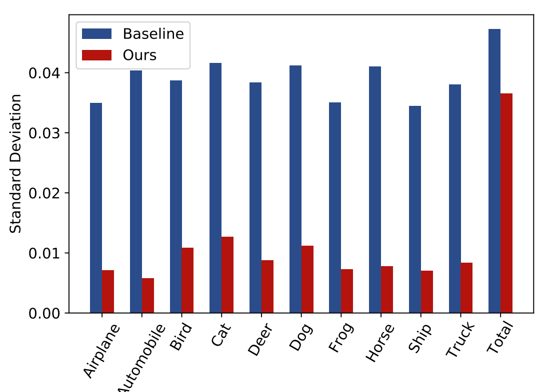


Figure 4. Compactness (*i.e.,* standard deviation) of the embedding clusters on CIFAR-10. Standard deviation is calculated on the embedding codes within the same annotation. The 'Total' is the compactness of the whole embedding codes in the test set.

Ablation Study. To validate that our method shrinks the regions of samples in the feature space, we also analyze the cluster compactness (*i.e.,* standard deviation of the cluster) of each class in the feature space on CIFAR-10, which is presented in Fig. 4. It can be observed that the class-wise standard deviation of our AdaptiveMix is much lower than that of the baseline. The entry of 'Total' measures the compactness of regions of all samples in the feature space. Fig. 4 shows that our method shrinks regions of samples in the feature space, compared with that of the baseline (*i.e.,* without AdaptiveMix).

Table 7. Accuracy (%) on CIFAR-10 based on WRN-28-10 trained with the various methods with orthogonal classifier (Orth.).

CIFAR10	FGSM (8/255)	PGD-8 (4/255)	PGD-16 (4/255)	CW-100 (c=0.01)	CW-100 (c=0.05)
Baseline	38.03	0.92	0.28	11.1	0.39
Mixup [54]	60.17	3.97	1.16	30.32	2.36
Orth. + Mixup	44.80	3.99	2.66	71.12	49.47
M.-Mixup [44]	59.32	7.97	2.97	51.47	11.12
Orth. + M.-Mixup	38.76	5.77	4.38	69.08	53.98
Ours	**74.18**	**32.12**	**22.12**	**81.39**	**74.72**

Table 8. Accuracy (%) on CIFAR-100 and Tiny-ImageNet against various adversarial attacks based on WRN-28-10 [53] and PreActResNet-18 [9] respectively.

Dataset	Method	FGSM (8/255)	PGD-8 (4/255)	PGD-16 (4/255)	CW-100 (c=0.01)	CW-100 (c=0.05)
C-100	Baseline	11.71	0.79	0.42	4.42	0.23
	Mixup [54]	27.34	0.28	0.11	4.83	0.28
	M.-Mixup [44]	**29.73**	1.19	0.49	10.75	0.77
	Ours	24.28	**8.22**	**7.40**	**42.02**	**26.18**
T-ImageNet	Baseline	4.26	0.81	0.60	27.92	7.52
	Mixup [54]	4.23	0.98	0.77	29.13	15.41
	M.-Mixup [44]	3.04	0.82	0.59	29.69	16.86
	Ours	**7.10**	**4.66**	**4.98**	**35.93**	**34.22**

Table 9. Accuracy (%) on CIFAR-10 based on WRN-28-10 trained with the proposed method under various noise terms σ.

Noise	FGSM (8/255)	PGD-8 (4/255)	PGD-16 (4/255)	CW-100 (c=0.01)	CW-100 (c=0.05)
σ=0.1	71.90	32.54	**23.31**	79.58	71.96
σ=0.01	71.56	34.04	25.96	80.68	72.00
σ=0.005	71.31	27.79	20.64	80.46	71.58
σ=0.001	70.51	27.42	17.98	79.47	67.00
σ=0.05	**74.18**	**32.12**	22.12	**81.39**	**74.72**

Robust Image Recognition. To evaluate the adversarial robustness of the proposed method, we compare it with Mixup [54], and Manifold Mixup [44] (denoted as M.-Mixup) and

Table 10. Accuracy (%) on CIFAR-10 based on WRN-28-10 trained with the proposed method using various of $alpha$ for the Beta distribution to generate mixing coefficient λ.

Alpha	FGSM (8/255)	PGD-8 (4/255)	PGD-16 (4/255)	CW-100 (c=0.01)	CW-100 (c=0.05)
α=2.0	71.27	31.85	**22.39**	79.18	70.50
α=1.0	**74.18**	**32.12**	22.12	**81.39**	**74.72**

Table 11. Accuracy (%) of the proposed AdaptiveMix on varying baselines and datasets. Res. stands for resolution of the input.

Dataset	Architecture	Res.	Baseline	Ours
CIFAR-10	WRN-28-10 [53]	32^2	96.11	**96.80**
CIFAR-100	WRN-28-10 [53]	32^2	80.82	**82.02**
T-ImageNet	PreActResNet-18 [9]	64^2	57.23	**60.59**
ImageNet	ResNet-50 [8]	128^2	67.38	**68.69**

present the results in Tab. 7 and Tab. 8. As listed in Tab. 7, the average classification accuracy of the proposed method can reach 56.91% on CIFAR-10, which surpasses Mixup and Manifold Mixup by margins of 37.31% and 30.34%, respectively. Furthermore, the proposed model is also tested on a large-scale dataset, *i.e.,* Tiny-ImageNet. As listed in Tab. 8, the proposed approach can achieve superior performance against different kinds of adversarial attacks, compared to the other interpolation-based methods. Concretely, by using our AdaptiveMix, the model can achieve an average accuracy of 17.38%, surpassing Manifold Mixup by an improvement of $\sim$7%.

Here, we analyze the influence caused by different values of hyper-parameters, including σ for the noise term, and α for Beta Distribution in AdaptiveMix. As listed in Tab. 9, noises on multiple levels σ=[0.1, 0.05, 0.01, 0.005, 0.001] are considered for the grid search. The best performance is achieved as σ=0.05. In terms of α, we conduct two settings for comparison, including α=1.0 and 2.0 in Tab. 10. The model trained with α=1.0 is observed to outperform the one with α=2.0 (*i.e.,* 56.91% *vs.* 55.04%).

Clean Image Recognition. To validate the effectiveness of the proposed method on image recognition, we test the proposed method on various standard datasets and compare the results compared with the baseline [8, 9, 53]. Tab. 11 shows our method improves the baseline on all the cases. In particular, on Tiny-ImageNet, 3% absolute improvement can be achieved by the proposed method. The experimental result indicates that AdaptiveMix also not only improves the robustness but also benefits the generalization.

OOD Detection. To validate the effectiveness of our method in OOD detection, we compare with state-of-the-art OOD detection approaches [32, 36, 49, 52] on various datasets. We refer to the accuracy of 1DS [52] as the upper bound of other methods in Tab. 12, since 1DS employs Monte Carlo (MC) sampling, which sacrifices computational efficiency for achieving high accuracy. Consequently, 1DS consumes a much higher time cost than other methods. We build a baseline named "1DS w/o MC" [52] removing MC sampling for 1DS, and our method combines "1DS w/o MC" with our AdaptiveMix. Tab. 12 shows the performance of "1DS w/o MC" is degraded due to the lack of MC sampling. However, our AdaptiveMix effectively improves "1DS w/o MC" without expensive computational cost.

Table 12. OOD detection on various OOD sets, where TIN-C, TIN-R, LSUN-C, and LSUN-R refer to the OOD set of Tiny ImageNet-Crop, Tiny ImageNet-Resize, LSUN-Crop, and LSUN-Resize, respectively. All values are F1 score (↑), † stands for the result reproduced by the open-source code.

ID Dataset	CIFAR10			
OOD Dataset	TIN-C	TIN-R	LSUN-C	LSUN-R
Methods using MC sampling				
1DS [52] (CVPR'21)	0.930	0.936	0.962	0.961
Methods which adopt OOD samples for validation and fine-tuning				
ODIN [26] (ICLR'18)	0.902	0.926	0.894	0.937
Mahalanobis [24] (NIPS'18)	0.985	0.969	0.985	0.975
Soft. Pred. [10] (ICLR'17)	0.803	0.807	0.794	0.815
Counterfactual [36] (ECCV'18)	0.636	0.635	0.650	0.648
CROSR [49] (CVPR'19)	0.733	0.763	0.714	0.731
OLTR [32] (CVPR'19)	0.860	0.852	0.877	0.877
1DS w/o MC † [52]	0.890	0.886	0.897	0.907
1DS w/o MC † +Ours	**0.922**	**0.911**	**0.934**	**0.937**

5. Conclusion

In this paper, we proposed a novel module named AdaptiveMix which is simple yet effectively improves the training of GANs. By reducing the distance between training samples and their convex combination in a dynamic manner, AdaptiveMix can shrink regions of training data in the feature space, enabling the stable training of GANs and improving the image quality of generated samples. We also demonstrate that AdaptiveMix is a reasonable way to ensure the approximate estimation of Lipschitz continuity. Besides image generation, we show that AdaptiveMix can be applied to other tasks, such as image classification and OOD detection, thanks to its plug- and-play property. Experimental results demonstrate that our method effectively improves the performance of baseline models on seven publicly available datasets with regard to various tasks.

Acknowledgement

This work was supported in part by the KAUST Office of Sponsored Research through the Visual Computing Center (VCC) funding, as well as, the SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence (SDAIA-KAUST AI); in part by the Key-Area Research and Development Program of Guangdong Province, China (No. 2018B010111001), and in part by National Key R&D Program of China (2018YFC2000702) and the Scientific and Technical Innovation 2030-"New Generation Artificial Intelligence" Project (No. 2020AAA0104100).

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223. PMLR, 2017. 1, 2, 4, 6

[2] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020. 5, 6

[3] Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In *International Conference on Machine Learning*, pages 854–863. PMLR, 2017. 4

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009. 3, 5

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27:2672–2680, 2014. 1, 6

[6] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. *arXiv preprint arXiv:1704.00028*, 2017. 2, 6

[7] Hongyu Guo, Yongyi Mao, and Richong Zhang. Mixup as locally linear out-of-manifold regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3714–3722, 2019. 2, 3

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 8

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016. 2, 7, 8

[10] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *International Conference on Learning Representations*, 2017. 8

[11] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2018. 2

[12] 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. *Advances in Neural Information Processing Systems*, 30, 2017. 1

[13] Minui Hong, Jinwoo Choi, and Gunhee Kim. Stylemix: Separating content and style for enhanced data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14862–14870, 2021. 3

[14] Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8710–8719, 2021. 2

[15] Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. Deceive d: Adaptive pseudo augmentation for gan training with limited data. *Advances in Neural Information Processing Systems*, 34:21655–21667, 2021. 2, 3, 6, 7

[16] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *International Conference on Learning Representations*, 2018. 1

[17] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, 33:12104–12114, 2020. 2, 3, 6, 7

[18] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021. 1

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 5, 6

[20] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020. 1, 2, 4, 6

[21] Jang-Hyun Kim, Wonho Choo, and Hyun Oh Song. Puzzle mix: Exploiting saliency and local statistics for optimal mixup. In *International Conference on Machine Learning*, pages 5275–5285. PMLR, 2020. 2

[22] Alex Krizhevsky. Learning multiple layers of features from tiny images. *Technical Report TR-2009, University of Toronto, Toronto*, 2009. 3, 5, 6

[23] Karol Kurach, Mario Lučić, Xiaohua Zhai, Marcin Michalski, and Sylvain Gelly. A large-scale study on regularization and normalization in gans. In *International conference on machine learning*, pages 3581–3590. PMLR, 2019. 1

[24] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in Neural Information Processing Systems*, 31, 2018. 3, 8

[25] Bing Li, Yuanlue Zhu, Yitong Wang, Chia-Wen Lin, Bernard Ghanem, and Linlin Shen. Anigan: Style-guided generative adversarial networks for unsupervised anime face generation. *IEEE Transactions on Multimedia*, 2021. 1

[26] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *International Conference on Learning Representations*, 2018. 3, 8

[27] Haozhe Liu, Bing Li, Haoqian Wu, Hanbang Liang, Yawen Huang, Yuexiang Li, Bernard Ghanem, and Yefeng Zheng. Combating mode collapse in gans via manifold entropy estimation. *arXiv preprint arXiv:2208.12055*, 2022. 1

[28] Haozhe Liu, Haoqian Wu, Weicheng Xie, Feng Liu, and Linlin Shen. Group-wise inhibition based feature regularization

for robust classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 478–486, 2021. 2

[29] Haozhe Liu, Wentian Zhang, Jinheng Xie, Haoqian Wu, Bing Li, Ziqi Zhang, Yuexiang Li, Yawen Huang, Bernard Ghanem, and Yefeng Zheng. Decoupled mixup for out-of-distribution visual recognition. In *European Conference on Computer Vision*, pages 451–464. Springer, 2023. 2

[30] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015. 5

[31] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 6

[32] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2537–2546, 2019. 3, 8

[33] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2794–2802, 2017. 6

[34] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018. 1, 2

[35] Aamir Mustafa and Rafał K Mantiuk. Transformation consistency regularization–a semi-supervised paradigm for image-to-image translation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 599–615. Springer, 2020. 3

[36] Lawrence Neal, Matthew Olson, Xiaoli Fern, Weng-Keen Wong, and Fuxin Li. Open set learning with counterfactual images. In *Proceedings of the European conference on computer vision*, pages 613–628, 2018. 8

[37] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 1, 2, 6

[38] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 5

[39] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in Neural Information Processing Systems*, 29, 2016. 1, 2, 6

[40] Jürgen Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proc. of the international conference on simulation of adaptive behavior: From animals to animats*, pages 222–227, 1991. 1

[41] Jürgen Schmidhuber. Generative adversarial networks are special cases of artificial curiosity (1990) and also closely related to predictability minimization (1991). *Neural Networks*, 127:58–66, 2020. 1

[42] Casper Kaae Sønderby, Jose Caballero, Lucas Theis, Wenzhe Shi, and Ferenc Huszár. Amortised map inference for image super-resolution. *International Conference on Learning Representations*, 2017. 6

[43] Hung-Yu Tseng, Lu Jiang, Ce Liu, Ming-Hsuan Yang, and Weilong Yang. Regularizing generative adversarial networks under limited data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7921–7931, 2021. 2, 6

[44] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, pages 6438–6447. PMLR, 2019. 2, 3, 4, 7

[45] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018. 1

[46] Zhengning Wang, Fanwei Zeng, Shuaicheng Liu, and Bing Zeng. Oaenet: Oriented attention ensemble for accurate facial expression recognition. *Pattern Recognition*, 112:107694, 2021. 3

[47] Yuanbo Xiangli, Yubin Deng, Bo Dai, Chen Change Loy, and Dahua Lin. Real or not real, that is the question. *International Conference on Learning Representations*, 2020. 6

[48] Jingkang Yang, Haoqi Wang, Litong Feng, Xiaopeng Yan, Huabin Zheng, Wayne Zhang, and Ziwei Liu. Semantically coherent out-of-distribution detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8301–8309, 2021. 3

[49] Ryota Yoshihashi, Wen Shao, Rei Kawakami, Shaodi You, Makoto Iida, and Takeshi Naemura. Classification-reconstruction learning for open-set recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4016–4025, 2019. 8

[50] Qing Yu and Kiyoharu Aizawa. Unsupervised out-of-distribution detection by maximum classifier discrepancy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9518–9526, 2019. 2

[51] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019. 2

[52] Alireza Zaeemzadeh, Niccolò Bisagno, Zeno Sambugaro, Nicola Conci, Nazanin Rahnavard, and Mubarak Shah. Out-of-distribution detection using union of 1-dimensional subspaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9452–9461, 2021. 2, 3, 5, 8

[53] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference*. British Machine Vision Association, 2016. 2, 7, 8

[54] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 2, 3, 5, 7

[55] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International Conference on Machine Learning*, pages 7354–7363. PMLR, 2019. 1

[56] Han Zhang, Zizhao Zhang, Augustus Odena, and Honglak Lee. Consistency regularization for generative adversarial networks. *International Conference on Learning Representations*, 2020. 2

[57] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3096–3105, 2019. 1

[58] Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. In *International Conference on Learning Representations*, 2017. 6

[59] Jiachen Zhong, Xuanqing Liu, and Cho-Jui Hsieh. Improving the speed and quality of gan by adversarial training. *arXiv preprint arXiv:2008.03364*, 2020. 2

[60] Mingchen Zhuge, Dehong Gao, Deng-Ping Fan, Linbo Jin, Ben Chen, Haoming Zhou, Minghui Qiu, and Ling Shao. Kaleido-bert: Vision-language pre-training on fashion domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12647–12657, June 2021. 3