**ABSTRACT** Generative Adversarial Networks (GANs) play the adversarial game between two neural networks: the generator and the discriminator. Many studies treat the discriminator’s outputs as an implicit posterior distribution prior to the input image distribution. Thus, increasing the discriminator’s output dimensions can represent richer information than a single output dimension of the discriminator. However, increasing the output dimensions will lead to a very strong discriminator, which can easily surpass the generator and break the balance of adversarial learning. Solving such conflict and elevating the generation quality of GANs remains challenging. Hence, we propose a simple yet effective method to solve this conflict problem based on a stochastic selecting method by extending the flipped and non-flipped non-saturating losses in BipGAN. We organized our experiments based on the famous BigGAN and StyleGAN models for comparison. Our experiments successfully validated our approach to strengthening the generation quality within limited output dimensions via several standard evaluation metrics and real-world datasets and achieved competitive results in the Human face generation task.

**INDEX TERMS** Stochastic discriminator, label flipping, image generation.

I. INTRODUCTION

Generative Adversarial Networks (GANs) [1] and its variants are one of the most successful generative models that maintain a dynamic balance between two neural networks: the generator and the discriminator. GANs can generate imitative images and defeat many other generative models, such as VAE [2], Flow [3], etc., owing to the discriminator and the adversarial training.

The discriminator in GANs could be treated as a binary classifier, which can recognize input images from the real image source or the generated fake image source. Meanwhile, some GANs’ studies (such as f-GAN, IPM-GANs, etc. [4], [5], [6], [7], [8], [9]) treated the discriminator’s outputs as an implicit posterior distribution, conditioning on parameters in the discriminator.

Thus, using multi-dimensional discriminator’s outputs rather than a single dimension is a helpful way to represent richer information. However, the generation quality cannot improve if the discriminator’s output dimensions are too high, as one ablation study in repulsive MMD-GAN [9]. Before reaching such high dimensions, richer information from the discriminator output can help to train a better generator. Then, if further increasing the discriminator’s output dimensions, the discriminator gradually surpasses the generator and thus breaks the balance of the adversarial learning and harms the generation quality. Besides, some famous GAN models [9], [10], [11] have to use multi-dimensional
discriminator’s outputs for specific purposes. For instance, PatchGAN quadratically increases the discriminator’s output dimensions, which may easily lead to an imbalance between the generator and the discriminator. Further research is necessary to avoid the conflict between higher dimensions and to keep the balance of adversarial learning.

BipGAN [12] explored another way to benefit GAN’s training based on multi-dimensional discriminator’s outputs. The authors set two outputs for the discriminator and then led outputs to two different optimization directions. Here, two types of outputs functionally work as two different kinds of discriminators: one for authenticity discrimination and the other for falsity discrimination. Such two types of discriminators in logic can work together to increase the posterior reliability [13] of the whole discriminator. Therefore, the adversarial training process benefits and improves the generation quality. BipGAN inspired us to solve the conflict problem mentioned before in the discriminator. We can use limited output dimensions to represent two distributions learned from the two types of outputs in BipGAN [12].

First, we set the output dimensions of the discriminator as $N$ pairs of outputs ($2N$ in all). One output is the authenticity loss in each pair, and the other is the falsity loss. Then, we randomly select part of the pairs only to back-propagate authenticity loss and the rest in the pair only for falsity loss. On this basis, our discriminator keeps double dimensions while holding the same volume of parameters as traditional GANs in the back-propagation to avoid the conflict problem.

All in all, we highlight our contributions to this work:

- We extend BipGAN to larger dimensional outputs without causing the conflict problem via our random selecting method;
- We tested the empirically best ratio and the selection policy between authenticity and falsity outputs;
- Our experiments validate the effectiveness of our approach on conditional GANs with multi-dimensional discriminator’s outputs based on several real-world datasets and achieve competitive generation quality in the Human face generation task.

II. BACKGROUND

The discriminator in GAN models was developed in many directions based on past research [9], [14], [15]. Increasing the output dimensions of the discriminator is a feasible and useful way to improve GAN’s training while harming its performance after reaching the summit at specific dimensions. In this section, we will introduce several past works based on multi-dimensional outputs in discriminators.

A. MULTI-DIMENSIONAL OUTPUTS IN THE DISCRIMINATOR

As validated in repulsive MMD-GAN [9], using multi-dimensional discriminator’s outputs can significantly improve the generation quality due to an information-rich representation of the input data. In contrast, the model generation ability faces the conflict problem when adding more output dimensions.

Maximum Mean Discrepancy (MMD) [16] loss function requires multi-dimensional discriminator’s outputs to accommodate enough information in Hilbert’s space. Repulsive MMD-GAN improved one successful past work: MMD-GAN [17], which utilized MMD as the loss metrics [11], [18] to minimize the distance between training data distribution and generated image distribution. Here is the original MMD loss [19], [20] in the discriminator step:

$$L_{MMD}^D = -\mathbb{E}_{x \sim P_{data}} [k(D(x), D(x'))] + 2 \times \mathbb{E}_{x \sim P_{data}, y \sim P_G} [k(D(x), D(y))] - \mathbb{E}_{y \sim P_G} [k(D(y), D(y'))].$$ (1)

Moreover, Eq. 2 is the repulsive MMD loss for the discriminator step:

$$L_{rep}^D = \mathbb{E}_{x \sim P_{data}} [k(D(x), D(x'))] - \mathbb{E}_{y \sim P_G} [k(D(y), D(y'))].$$ (2)

where $k(D(a), D(b))$ is the kernel embedding function (e.g., RBF kernel) and $(a, b)$ could be the multi-dimensional discriminator outputs $x$ or $y$ as defined in [9]. Such function can kernelize $x$ or $y$ from the $N$ dimensional vector to a $N^2$ matrix. $P_{data}$ and $P_G$ are the real image distribution and the generated image distribution, respectively (the same below).

The authors showed that the generation quality is increasing from 1 to 16 dimensions in terms of Inception Score (IS) [21], and Fréchet Inception Distance (FID) [22] can engage a better performance until 64 dimensions. Both FID and IS are decreased when the discriminator’s output dimensions achieve 256.

B. BipGAN

BipGAN [12] creates a novel way to improve the generator with two-dimensional outputs in the discriminator. BipGAN’s discriminator guides one output to one learning direction as traditional GANs and guides the other one to an opposite direction. The authors name two outputs and losses as authenticity or falsity. Hence one can evaluate the possibility of the input source, which belongs to the real image distribution and vice versa to the fake image distribution. Therefore, leading the output distribution in one direction can ensure the discriminative confidence of either real distribution or generated distribution, yet hard to guide both distributions simultaneously. Hence, one natural idea is to separate the discriminative task into two discriminator outputs: one for real distribution (authenticity loss $L_A$) and the other one for generated distribution (falsity loss $L_F$). Such two outputs can offer two types of discrimination and work together to promote the posterior confidence of the whole discriminator.

$$L_A = \mathbb{E}_{x \sim P_{data}} \log (D_A(x)) + \mathbb{E}_{y \sim P_G} \log (1 - D_A(y)),$$ (3)

$$L_F = \mathbb{E}_{x \sim P_{data}} \log (1 - D_F(x)) + \mathbb{E}_{y \sim P_G} \log (D_F(y)).$$ (4)
where $D_A$ refers to the discriminator with authenticity loss output and $D_F$ for falsity loss output. By optimizing the sum of $L_A$ and $L_F$ as the discriminator loss, the discriminator can receive information from different outputs, thus enhancing adversarial learning.

### C. NON-SATURATING LOSS AND LABEL FLIPPING

Here we give an extended analysis of authenticity loss from $D_A$ and falsity loss from $D_F$. As explained in [1] and [23], two types of generator losses refer to two kinds of adversarial games: the mini-max game and the heuristic non-saturating game.

\[
L_{\text{saturating}} = \mathbb{E}_{\hat{y} \sim P_{\hat{y}}} \left[ \log (1 - D(y)) \right],
\]
\[
L_{\text{non-saturating}} = \mathbb{E}_{\hat{y} \sim P_{\hat{y}}} \left[ \log (D(y)) \right].
\]

(Eq. 5)

In Eq. 5, saturating generator loss indicates the mini-max game, and non-saturating generator loss indicates the heuristic non-saturating game.

One past work [21], the author’s talk\(^1\) and the related practical project [24] mentioned that: flipping target labels (set real=0 and fake=1 in the discriminator loss and fake=0 in the generator loss) can obtain empirically better results than results from the non-flipped version of the non-saturating game.

Here, we explain that the saturating generator loss can work as a non-saturating loss under the flipped targets. As shown in Fig. 1, the authenticity loss in BipGAN [12] is exactly the same as in this non-saturating game, and this state (before label-flipping) used the $D_A$ (authenticity discriminator). First we show two curves for $J^G = -\log(D(G(z)))$ (green curve) and $J^G = \log(1 - D(G(z)))$ (blue curve) as showed in [23]. The author in [23] answered why the blue curve is saturating: since the discriminator can easily recognize real and fake images at initial training steps, $D_A(G(z))$ is almost 0 at the beginning. Hence, $J^G = \log(1 - D_A(G(z)))$ is almost 0 and hard to supply effective gradients for learning (blue arrow). As one solution, using $J^G = -\log(D_A(G(z)))$ can improve the learning efficiency at initial steps (green arrow). Then, if we flip the training target (which is the same as the falsity output’s target in [12]), $D_F(G(z))$ is almost 1 at the beginning. Thus, we can ascend gradients from $J^G = \log(1 - D_F(G(z)))$ to avoid being saturating (red arrow), which is originally saturated in the non-flipped version. As shown in different arrows, the initial value of $D(G(z))$ matters the criterion value and thus affects gradients and the training effectiveness for the back-propagation steps.

Besides, the label flipping trick could be unified in a single value function (the negative version of $V(G,D)$ in [1]) under Jensen–Shannon divergence [1], [23]. Therefore, the label flipping trick can inherit theoretical conclusions from the mini-max game in [1], yet has not been proved in the heuristic non-saturating game as introduced in [8] and [25].

All related literature point to the fact that flipping the discriminator’s target is a useful way to enhance the GANs’ training, and how to fully use such a mechanism remains to be studied.

### D. VECTOR-BASED DISCRIMINATOR

Here we define the discriminator with multi-dimensional outputs (use a linear layer to map the last feature map to a multi-dimensional vector) as the vector-based discriminator. Some past works [14], [15] and our preliminary experiments in Tab. 1 validated the empirical advantages (before reaching its summit performance) of using the vector-based discriminator, yet not determined its reason. Reference [26] proved that vector-based discriminator could acquire richer information.

---

\(^1\)“How to train a GAN”, Soumith Chintala, NeurIPS 2016 Workshop on Adversarial Training.

\(^2\)https://www.youtube.com/watch?v=myGAju4L7O8

---

**FIGURE 1.** Saturating and non-saturating criterion curves in the mini-max game and the heuristic non-saturating game. Arrows indicate their initial training states. Details are in Section II-C.
under Wasserstein distance. In this work, we show the corollary from [26] to prove a similar conclusion while under the f-divergence [4] for our work.

Corollary 1: Define f-divergence between two distributions \( P \) and \( Q \) as \( \text{Div}_{f}^{(m)}(P||Q) \) based on the m-dimensional outputs discriminator. If two positive integer \( n < n' \), then for any distributions \( P, Q \in P(\mathbb{R}) \), we have \( \text{Div}_{f}^{(m)}(P||Q) \leq \text{Div}_{f}^{(n)}(P||Q) \).

Proof: Divide can be transferred to the problem that maximizes the discrepancy between two expectations based on Fenchel conjugate [27]:

\[
\text{Div}_{f}^{(m)}(P||Q) = \sup_{T^{(m)} \in \Gamma} \left( E_{x \sim P}[f^{(m)}(x)] - E_{x \sim Q}[f^{*}(T^{(m)}(x))] \right),
\]

where \( f \) is a convex function which is used to define the f-divergence as in [4] and \( f^{*} \) is its conjugate function as defined in [27] and [4]. \( \Gamma \) is the arbitrary class of functions \( T^{(m)} \). Hence larger \( m \) can offer more candidate functions in \( \Gamma \) when finding the optimal \( T^{(m)} \) for achieving the maximal discrepancy, thus obtaining the proof. \( \square \)

E. PATCH-BASED DISCRIMINATOR

Patch-based discriminator worked as a vital part in pix2pix [10]. GANs can generate vivid and high-quality images after remarkable improvements in training stability and network architectures. At the same time, high-resolution images can bring in extra instability and learning hardness in GANs’ training. Some works tried to solve this problem based on progressive training policy or create swollen networks, while one effective way is the Patch-based discriminator. Dividing a large image into several small patches in GANs’ training can significantly increase each patch’s generation quality and diversity.

Based on past works’ conclusions, patch discriminator can hardly obtain performance improvement via increasing output dimensions because the number of its output dimensions is already higher than others. Either quadratically increasing the number of patches or duplicating a curtein number of patches in a third dimension may easily cause the conflict problem. Therefore, one feasible way to acquire both advantages from the multi-dimensional discriminator’s outputs and the patch discriminator is to apply our stochastic discriminator to the patch discriminator.

III. STOCHASTIC OUTPUTS IN THE DISCRIMINATOR

A. MOTIVATION: CONFLICT PROBLEM WHEN INCREASING OUTPUT DIMENSIONS

As validated in repulsive MMD-GAN [9], increasing output dimensions of the discriminator can obtain significant performance improvement at the beginning while facing damage when dimensions become higher. After reaching such empirically best dimensions, the conflict problem happens when keeping on increasing the output dimensions. This problem happens in the MMD metric, as well as GANs with the non-saturating loss. Hence, how to break through such a problem restrict the model performance.

B. METHOD: RANDOMLY SELECTING OUTPUTS FOR BACK-PROPAGATION

To overcome the conflict problem brought by the higher dimensions in the discriminator outputs, we propose to modify BipGAN’s bipolar discriminator to a randomly selected discriminator. In BipGAN, the authors combined non-saturating loss and flipped non-saturating loss to obtain better posterior confidence in the discriminator. Such a posterior discriminator gains advantages from two types of losses. Naturally, we can increase discriminator’s two outputs to higher dimensions: a half for authenticity loss and a half for falsity loss. Suppose we randomly abandon some authenticity and rest for falsity outputs. In that case, the practical 2N dimensions become N. Based on conclusions from repulsive MMD-GAN and our preliminary experiments, the reduction can help the discriminator overcome the conflict problem. Thus, we generalize our method to patch-based discriminators. Duplicating all patches’ outputs in the discriminator or quadratically increasing the number of patches can easily trigger the conflict problem. In a word, our random selection method can help multiple-output-based discriminators eliminate restrictions.

In Fig. 2, the discriminator outputs are shaped as \( (2N^2) \) while patch-based outputs in red are for authenticity loss and blue for falsity loss. The significant difference between this work and BipGAN is that the BipGAN method will use all red and blue patch-based outputs for the back-propagation. At the same time, ours randomly selects part of the red outputs from authenticity outputs. The remaining parts (in shallow red) will be neglected, and we enable outputs from falsity outputs (in deep blue). Finally, patch-based outputs from two losses (deep red and blue) are concatenated together as the stochastic output loss metrics with the shape of \( (N^2) \) for the backward step. We randomly select outputs for backward. Thus, the mixing method and ratio of authenticity and falsity outputs are significant, and we will explore these issues in our experimental ablations.

C. ADVERSARIAL LOSS FUNCTION

Our method extends BipGAN’s generator loss function [12] to n-dimensional cases:

\[
L_G = E_{y \sim P_G} \left[ \log(D_n(y)) \right] + E_{y \sim P_G} \left[ \log(1 - D_n(y)) \right].
\]

Our discriminator loss function is based on Eq. 3 and Eq. 4, with the random vector mask \( k \) to control the back-propagation and extend them to n-dimensions:

\[
L_D = k \odot L_n^{(n)} + (1 - k) \odot L_n^{(n)}, \quad k \sim B(1, p).
\]

In Algorithm 1, we show how our method works in a patch-based discriminator.

VOLUME 10, 2022
We organized all experiments based on one common and powerful architecture, namely BigGAN [31] for the experimental comparisons. BigGAN is concise in losses and easy to train compared with other state-of-the-art models. In order to avoid uncertainty via other factors (such as the penalty, other loss metrics, etc.), we chose BigGAN and edited the discriminator output layer of BigGAN merely. We removed the final pooling layer in patch-based experiments to create receptive fields for input patches. We used the default hyper-parameters as in BigGAN experiments, and all models were trained by Adam optimizer [32]. The exponential

![Algorithm 1 Our Method via the Randomly Selecting Training Strategy in the Discriminator With Multi-Dimensional Outputs](image)

**IV. EXPERIMENTS**

**A. SETTINGS**

We set original non-saturating loss in vanilla-GAN [1], flipped non-saturating loss in [21] and acquired both losses in BipGAN [12] as baseline methods to compare to our work. Besides, as we used the multiple outputs discriminator like repulsive MMD-GAN, we also set it as another baseline (mixed RBF kernel w/ \( \{1, \sqrt{2}, 2, 2\sqrt{2}, 4\} \)). Firstly, we validate the empirically best dimensional number based on a set of ablation studies. Next, we subdivide stochastic methods as ‘Hard-mix’ and ‘Soft-mix’ and test their performance. Afterward, we visualize the trend chart of such ablations. Finally, we test our methods and compare them to three past works based on several real-world datasets.

1) **DATASETS**

We tested and compared our methods in our experiments with past works based on four widely used real-world datasets. CIFAR-10 [28] (50k training images, 10 classes, 32 × 32 resolution) and CIFAR-100 [28] (50k training images, 100 classes, 32 × 32 resolution) are two common standard thumbnail real-world image datasets. We organized ablation experiments via CIFAR-10 to test empirically best parameters.

Tiny-ImageNet [29] (100k training images, 200 classes, 64 × 64 resolution) is a replacement dataset for the standard 1k classes ImageNet dataset. Training standard is time and electricity expensive and normally costs more than two weeks for a single model. Therefore we used Tiny-ImageNet to test and compare our methods with past works.

CelebA [30] (203k training images, 64 × 64 resolution) is a famous human face dataset. After alignment and cropping, each image has a 64 × 64 resolution. Compared with CIFAR and Tiny-ImageNet datasets, CelebA only contains human face images and no other categories.

2) **NETWORK ARCHITECTURE**

We organized all experiments based on one common and powerful architecture, namely BigGAN [31] for the experimental comparisons. BigGAN is concise in losses and easy to train compared with other state-of-the-art models. In order to avoid uncertainty via other factors (such as the penalty, other loss metrics, etc.), we chose BigGAN and edited the discriminator output layer of BigGAN merely. We removed the final pooling layer in patch-based experiments to create receptive fields for input patches. We used the default hyper-parameters as in BigGAN experiments, and all models were trained by Adam optimizer [32]. The exponential
moving average (EMA) was started after the first 1k training iterations. We trained each experiment 10k iterations and reported its best FID and corresponding IS. The batch size was 64 for Tiny-ImageNet and CelebA experiments and 256 for CIFAR experiments. The dimension number of the input noise is 128 for all experiments. We also test the unconditional generation tasks based on StyleGAN2 [34], [35], one famous architecture in StyleGAN series [36], which can reflect the performance of each loss metric.

3) EVALUATION METRICS
We used Fréchet Inception Distance [22] (FID, lower is better) and Inception Score [21] (IS, higher is better) as the evaluation methods. We calculated and recorded the IS and FID score every 1k training steps and sampled 50k random samples to get the FID and IS. IS compares the feature-level similarity between sampled images and pretrained ImageNet [37]. Therefore, for non-imagenet-like datasets, higher IS can reflect a higher similarity to the ImageNet dataset. FID can usually reflect the generation quality better when using both generated and training samples [21], [38].

B. PRELIMINARY EXPERIMENT: DIMENSIONS
As validated in repulsive MMD-GAN [9], the best generation quality is obtained in 16 dimensions for IS and 64 dimensions for FID. On the contrary, FID and IS worsen when output dimensions achieve 256. In our experiments, we turn to use the powerful BigGAN model, and thus the training situation may change in our work. Therefore, we use the CIFAR-10 dataset to train several ablation studies to test our experiments’ empirically best output dimensions.

We test two cases that indicate two types of outputs. The trivial vector-based outputs are the duplicated version of a traditional discriminator with a single output. We duplicate the output from one dimension to \{2, 8, 16, 64, 256\} times.

From Tab. 1 we find that the generation quality is increased until eight dimensions and causes the conflict problem after 16 dimensions in terms of FID and IS. Regarding the trend, we plot FID scores within output dimensions in Fig. 3(a).

The other one is the patch-based outputs in the discriminator. In our ablations, we separate the input image to the 8 × 8 patches. Therefore, the de facto number of output dimensions is 64. We also duplicated such 64 patches four times and obtained 256 dimensions. On the contrary, further duplication cannot improve the generation quality in FID and IS, as shown in Tab. 1.

C. PRELIMINARY EXPERIMENT: MIXING RATIO
Our stochastic methods mix authenticity outputs and falsity outputs to acquire both advantages without a double size of parameters in the last layer at back-propagation. Hence, how to mix such two types of outputs worth the discussion. We control the ratio between authenticity outputs and falsity outputs every ten percent, namely ‘Hard-mix’. For instance, if we set the ratio between authenticity outputs and falsity outputs as 10%/90%, it means that about 10% of authenticity outputs will be selected under Binomial distribution. Besides, to determine whether reducing half of the parameters in a backward step can avoid the conflict problem, we also set a comparison group, namely ‘Soft-mix’. ‘Soft-mix’ group experiments applied the Uniform distribution \( U (0, p') \) onto each specific output, similar to \( k \sim B (1, p) \) in Eq. 8. The \( p' \) control the maximum ratio of authenticity outputs as in Eq. 9.

\[
L \_D = k' \odot L \_A \_n + (1 - k') \odot L \_F \_n, \quad k' \sim U (0, p'). \tag{9}
\]

In this case, both authenticity outputs and falsity outputs can back-propagate gradients under the specific ratios \( k' \).
As shown in Tab. 2, generation quality performances in ‘Soft-mix’ group are generally poorer than ‘Hard-mix’ group. Such results indicate that more outputs can cause the conflict problem and make it hard to lead to ideal performance. Besides, we plot the FID trend curve accomplished with the ratio variety in Fig. 3(b,c). Both hard and soft versions demonstrate a similar phenomenon. Once the ratio becomes 20% for either authenticity outputs or falsity outputs, the generation quality gains its advantages. Thus, both Fig. 3(b) and (c) demonstrate a W-shaped trend. Based on such ablations, we tend to test three combinations in our main experiments: {20%/80%, 50%/50%, 80%/20%}.

### D. RESULTS ANALYSIS

As shown in Tab. 3, we compare our methods to three past methods, which are Vanilla GAN with non-saturating loss, flipped non-saturating loss, and BipGAN. We set four different settings for our methods. As shown in Tab. 2, soft-mix methods cannot obtain their advantages over hard-mix methods. Thus, we show the “50%/50% soft-mix” setting as the lower bound of our method. Similarly, “50%/50% hard-mix” setting is selected as another comparison. Due to the outstanding performances of “20%/80%” and its reverse case in hard-mix settings, we use them to demonstrate the upper bound performances of our methods. Besides, we repeat all settings from the vector-based discriminator outputs to the patch-based discriminator outputs. The vector-based outputs setting is the unidirectional expansion of the traditional single output discriminator in the output layer. The discriminator can use multiple dimensions to represent real or fake image distributions in this case. The patch-based outputs setting separates input images into small patches.

In CIFAR-10 and CIFAR-100 experiments, 1) Our methods can achieve superior performances than past works under 20% and 80% combinations in multi-categories datasets both in Inception Score and FID evaluations; 2) Changing output ratios between authenticity outputs and falsity outputs can obviously affect generation qualities; 3) Generally, the “50%/50%” combination in both soft and hard mixing methods performed worse results.

In CelebA experiments, our methods perform different results on vector-based and patch-based results. In vector-based experiments, the discriminator’s output dimensions are lower than the patch-based ones. Thus, patch-based discriminators are easier to surpass the generators. Our “20%/80%” method demonstrates its advantage against LSGAN losses and WassersteinGAN losses in such cases. In Table 2, both non-saturating and flipped non-saturating losses cannot defeat our hard-mixing methods in some ratios for CIFAR-10. Here, flipped non-saturating losses’ performances are better in FID. All in all, other losses can obtain their advantages in specific datasets or settings, while ours are generally better choices in most cases.

Based on the analysis of real-world datasets, we found that our method supplied the possibility to change the ratio parameter for achieving superior model performance. The more complex datasets (more categories) can obtain a more significant improvement. Therefore, we tested a more complex real-world dataset to validate our hypothesis.

In Tiny-ImageNet experiments, the number of categories increased to 200. With the increased learning hardness, our methods demonstrated their advantages against past methods. In Tab. 5, the mixing ratio showed similar phenomena to other experiments: “50%/50%” case showed even worse results than past methods. At the same time, 20% of authenticity loss with 80% of falsity loss obtained the best performances both in Inception Score and FID score. Furthermore, the repulsive-MMD method performed quite unstable among different datasets, which are generally better in complex cases. We guess that MMD-based losses are sensitive to the kernel parameters, while we used the default settings.

Finally, we doubt the model effect because of the different performances in conditional (CIFAR, Tiny-ImageNet) and unimodal (CelebA) training tasks. Therefore, we tested another network architecture with the unconditional training task, namely StyleGAN2 [34], [35]. In Tab. 6, the unconditional CIFAR-10 generation results were similar to the cases in vector-based BigGAN experiments. Repulsive-MMD losses and BipGAN losses are outstanding in IS, while our “80%/20%” hard mixing method can always gain its advantage. Thus, we can conclude that different loss metrics are more sensitive to output dimensionality and data resolution than class numbers and conditional labels.

Besides, we also show some samples in Fig. 4 and Fig. 5 from our patch-based experiments. All methods can avoid mode collapse when generating images. Generally, our methods can generate more details in some cases or samples and thus reflect higher evaluation scores.

### V. RELATED WORKS

#### A. D2GAN

D2GAN [39] proposed a two-discriminator-based method to stabilize GANs’ training. The significant difference between our work and D2GAN includes two aspects. First, D2GAN has two discriminators, and ours only have one physical discriminator. In BipGAN and this work, though different output dimensions indicate the authenticity of the falsity, the
TABLE 3. Vector-based discriminator experimental results. Here we show the best FID (lower is better) and their corresponding IS (higher is better) during training processes. As IS is meaningless for the human face dataset, we only list FID for CelebA experiments.

| Method                  | CIFAR-10 | CIFAR-100 | CelebA |
|-------------------------|----------|-----------|--------|
| Non-saturating          | 7.799±0.107 | 9.6245 | 9.080±0.137 | 15.6537 | 13.0888 |
| Flipped Non-saturating  | 7.805±0.114 | 9.3634 | 8.205±0.109 | 12.9514 | 9.3393 |
| BIPGAN                  | 8.144±0.094 | **7.0244** | 8.642±0.101 | 13.5311 | 12.1057 |
| Repulsive-MMD           | **8.569 ± 0.113** | 7.4262 | 6.444±0.074 | 27.7473 | 12.7859 |
| LSGAN                   | 8.348±0.095 | 7.5584 | 9.146±0.129 | **10.5529** | 7.4517 |
| WassersteinGAN          | 7.554±0.086 | 10.9862 | **9.212 ± 0.151** | 12.0728 | **7.1389** |
| Soft-Mix(50%/50%)       | 7.003±0.100 | 14.7598 | 7.144±0.093 | 17.4078 | 10.8762 |
| Hard-Mix(80%/20%)       | **8.333 ± 0.095** | **5.6735** | **9.249 ± 0.174** | **10.4716** | 10.3713 |
| Hard-Mix(50%/50%)       | 8.012±0.142 | 8.8731 | 8.191±0.115 | 12.5690 | 10.3493 |
| Hard-Mix(20%/80%)       | 8.232±0.132 | 6.4865 | 7.767±0.113 | 14.9442 | **10.0119** |

TABLE 4. Patch-based discriminator experimental results.

| Method                  | CIFAR-10 | CIFAR-100 | CelebA |
|-------------------------|----------|-----------|--------|
| Non-saturating          | 8.195±0.091 | 6.9595 | 8.572±0.107 | 14.2093 | 8.3469 |
| Flipped non-saturating  | 8.204±0.161 | 7.6177 | 8.231±0.152 | 14.6047 | **6.1441** |
| BIPGAN                  | 8.153±0.077 | 7.0574 | 8.517±0.109 | 14.1916 | 7.8056 |
| Repulsive-MMD           | 8.171±0.056 | 9.9782 | **8.715 ± 0.142** | **11.4868** | 9.5403 |
| LSGAN                   | 7.992±0.137 | 12.2579 | 8.507±0.161 | 17.6681 | 8.5577 |
| WassersteinGAN          | **8.272 ± 0.084** | **6.7924** | 8.310±0.144 | 25.7594 | 8.7840 |
| Soft-Mix(50%/50%)       | 8.251±0.108 | 10.5254 | 7.981±0.095 | 12.6615 | 8.7079 |
| Hard-Mix(80%/20%)       | 8.243±0.078 | 6.8911 | 9.117±0.142 | 11.9542 | 7.7301 |
| Hard-Mix(50%/50%)       | 7.966±0.068 | 7.8932 | 8.170±0.103 | 14.8969 | 7.1101 |
| Hard-Mix(20%/80%)       | **8.520 ± 0.106** | **6.7397** | **9.497 ± 0.091** | **7.4497** | **6.7071** |

TABLE 5. Tiny-ImageNet results on patch-based GANs.

| Method                  | Tiny-ImageNet:200 |
|-------------------------|--------------------|
| Non-saturating          | 10.411±0.112 | 37.7696 |
| Flipped non-saturating  | 8.272±0.084 | 47.8994 |
| BIPGAN                  | 10.378±0.171 | 36.8144 |
| Repulsive-MMD           | 10.919±0.210 | **25.9641** |
| LSGAN                   | **11.654 ± 0.191** | 27.5168 |
| WassersteinGAN          | 9.962±0.135 | 44.2395 |
| Soft-Mix(50%/50%)       | 10.107±0.191 | 32.4059 |
| Hard-Mix(80%/20%)       | 10.763±0.161 | 28.9783 |
| Hard-Mix(50%/50%)       | 8.913±0.071 | 45.709 |
| Hard-Mix(20%/80%)       | **11.678 ± 0.237** | **25.1708** |

TABLE 6. CIFAR-10 results on the vector-based StyleGAN.

| Method                  | CIFAR-10 |
|-------------------------|----------|
| Non-saturating          | 7.636±0.060 | 19.5916 |
| Flipped non-saturating  | 8.184±0.121 | 17.7838 |
| BIPGAN                  | 8.103±0.061 | 18.4591 |
| Repulsive-MMD           | **8.425 ± 0.084** | **17.3505** |
| LSGAN                   | 7.236±0.095 | 24.9832 |
| WassersteinGAN          | 7.580±0.064 | 20.3673 |
| Soft-Mix(50%/50%)       | 7.836±0.096 | 21.4237 |
| Hard-Mix(80%/20%)       | **8.174 ± 0.048** | **15.7352** |
| Hard-Mix(50%/50%)       | 7.679±0.119 | 20.7171 |
| Hard-Mix(20%/80%)       | 7.527±0.0470 | 17.6184 |

feature extracting function can be solved based on other layers. In contrast, D2GAN needs two discriminators to keep the training stability accomplished with their dual loss functions. Second, our work is based on a stochastic selecting policy, while D2GAN is similar to our soft-mix experiments with a fixed ratio. Our experiments showed that hard-mix methods obtained better results due to fewer dimensions by half.
B. DROPOUT

Dropout [40] is a well-known technique to improve neural networks. Randomly masking some fully connected neural in each training iteration can solve or improve some issues. Here in our work, we randomly select the discriminator outputs, similar to operating Dropout onto the last layer. There are two significant differences: one is that Dropout neglects the selected ones while ours will use the other loss to replace the selected ones. Second, many past works use Dropout in all layers; hence the gradient vanishing problems happen in different layers. In contrast, we only change the last layer because the two losses can share the feature extractor as one physical discriminator to reduce the computational cost.

FIGURE 4. Generated CIFAR samples in patch-based experiments. † indicated our stochastically flipped method.
C. MIXUP

Mixup [41] is a useful way to augment data by linearly mixing input images and output labels. In our ‘soft-mix’ experiments, we only mixed the discriminator outputs. These outputs are extracted from the posterior distribution, which cannot refer to the linear interpolation in the original space.

VI. CONCLUSION AND DISCUSSION

This work extended an improved way to fully use the advantages of the flipped and non-flipped losses to supply two types of output distributions, which can offer richer information without a double size of output dimensions during the backpropagation. Both flipped and non-flipped losses have the equivalent topology strength [42] and thus can work together. Other types of loss metrics may not be as convenient as the binary cross entropy loss. For instance, reversely using the MMD loss may cause the attractive problem as in [9]. If we find more useful loss metrics, the discriminator’s outputs could be used to represent more information within limited output dimensions.

In the future, we would like to investigate a non-parametric way to determine the optimal ratio between different loss metrics. Besides, developing f-divergence [4] families or IPM [5], [6], [7], [8], [9] families may offer more candidate metrics for our stochastic discriminator.

ACKNOWLEDGMENT

Part of the research results was obtained from the commissioned research (No. 225) by National Institute of Information and Communications Technology (NICT), Japan.

REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Proc. Adv. Neural Inf. Process. Syst., vol. 27, 2014.
[2] D. P. Kingma and M. Welling, “Auto-encoding variational Bayes,” 2013, arXiv:1312.6114.
[3] D. Rezende and S. Mohamed, “Variational inference with normalizing flows,” in Proc. Int. Conf. Mach. Learn., 2015, pp. 1530–1538.
[4] S. Nowozin, B. Cseke, and R. Tomioka, “f-GAN: Training generative neural samplers using variational divergence minimization,” in Proc. Adv. Neural Inf. Process. Syst., vol. 29, 2016, pp. 1–9.
[5] Y. Mroueh, C.-L. Li, T. Sercu, A. Raj, and Y. Cheng, “Sobolev GAN,” in Proc. Int. Conf. Learn. Represent., 2018.
[6] Y. Mroueh and T. Sercu, “Fisher GAN,” in Proc. Adv. Neural Inf. Process. Syst., vol. 30, 2017.
[7] Y. Mroueh, T. Sercu, and V. Goel, “MCGAN: Mean and covariance feature matching GAN,” in Proc. Int. Conf. Mach. Learn., 2017, pp. 2527–2535.
[8] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein generative adversarial networks,” in Proc. Int. Conf. Mach. Learn., 2017, pp. 214–223.
[9] W. Wang, Y. Sun, and S. Halgamuge, “Improving MMD-GAN training with repulsive loss function,” in Proc. Int. Conf. Learn. Represent., 2019.
[10] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in Proc. IEEE Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 1125–1134.
[11] C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Póczos, “MMD GAN: Towards deeper understanding of moment matching network,” in Proc. Adv. Neural Inf. Process. Syst., vol. 30, 2017.
[12] R. Yang and H. Nakayama, “Bipolar GAN: Double check the solution space and lighten false positive errors in generative adversarial nets,” in Proc. IEEE Int. Conf. Image Process. (ICIP), Sep. 2019, pp. 4260–4264.
[13] J. A. Hartigan, Bayes Theory, Cham, Switzerland: Springer, 2012.
[14] S. W. Park and J. Kwon, “Sphere generative adversarial network based on geometric moment matching,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 4292–4301.
[15] A. Semenov, V. Boginski, and E. L. Pasiliao, “Neural networks with multidimensional cross-entropy loss functions,” in Proc. Int. Conf. Comput. Data Social Netw. Cham, Switzerland: Springer, 2019, pp. 57–62.
RUI YANG received the M.S. degree from the University of Tokyo, in 2018. He is currently pursuing the doctoral degree with the Nakayama Laboratory, Graduate School of Information Science and Technology, The University of Tokyo. His research interests include generative adversarial networks, image generation, and deep learning.

DUC MINH VO received the M.S. degree in computer science from The University of Science, Vietnam National University Ho Chi Minh City, Vietnam, in 2016, and the Ph.D. degree in computer science from The Graduate University for Advanced Studies (SOKENDAI) in alliance with National Institute of Informatics (NII), Japan, in 2020. He is currently a Postdoctoral Researcher with The University of Tokyo. His research interests include image recognition and generation and language and vision. He is a regular reviewer of international conferences in computer vision.

HIDEKI NAKAYAMA (Member, IEEE) received the master’s and Ph.D. degrees in information science from the University of Tokyo, Japan, in 2008 and 2011, respectively. From 2012 to 2018, he was an Assistant Professor at the Graduate School of Information Science and Technology, The University of Tokyo, where he has been an Associate Professor, since April 2018. He is also a Faculty Member with the International Research Center for Neurointelligence (IRCN) and a Visiting Researcher with the National Institute of Advanced Industrial Science and Technology (AIST). His research interests include generic image recognition, natural language processing, and deep learning.

**REFERENCES**

[16] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, “A kernel two-sample test,” J. Mach. Learn. Res., vol. 13, no. 1, pp. 723–773, Jan. 2012.

[17] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, “Demystifying MMD GANs,” 2018, arXiv:1801.04101.

[18] T. Unterthiner, B. Nessler, C. Seward, G. Klambauer, M. Heusel, H. Ramsauer, and S. Hochreiter, “Coulomb GANs: Provably optimal Nash equilibrium via potential fields,” 2017, arXiv:1708.08819.

[19] Y. Li, K. Swersky, and R. Zemel, “Generative moment matching networks,” in Proc. Int. Conf. Mach. Learn., 2015, pp. 1718–1727.

[20] G. K. Dziugaite, D. M. Roy, and Z. Ghahramani, “Training generative neural networks via maximum mean discrepancy optimization,” 2015, arXiv:1505.03906.

[21] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved techniques for training GANs,” in Proc. Adv. Neural Inf. Process. Syst., vol. 29, 2016.

[22] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “GANs trained by a two-time-scale update rule converge to a local Nash equilibrium,” in Proc. Adv. Neural Inf. Process. Syst., vol. 30, 2017.

[23] I. Goodfellow, “NIPS 2016 tutorial: Generative adversarial networks,” 2017, arXiv:1701.00160.

[24] S. Chintala, E. Denton, M. Arjovsky, and M. Mathieu, How to Train a GAN? Tips and Tricks to Make GANs Work. San Francisco, CA, USA: GitHub, Dec. 2016.

[25] M. Arjovsky and L. Bottou, “Towards principled methods for training generative adversarial networks,” 2017, arXiv:1701.04862.

[26] M. Dai, H. Hang, and A. Srivastava, “Rethinking multidimensional discriminator output for generative adversarial networks,” 2021, arXiv:2109.03378.

[27] J.-B. Hiriart-Urruty and C. Lemaréchal, Fundamentals of Convex Analysis. Cham, Switzerland: Springer, 2012.

[28] A. Krizhevsky, “Learning multiple layers of features from tiny images,” M.S. thesis, Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, 2009.

[29] Y. Le and X. Yang, “Tiny ImageNet visual recognition challenge,” CS 231N, vol. 7, no. 7, p. 3, 2015.

[30] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 3730–3738.

[31] A. Brock, J. Donahue, and K. Simonyan, “Large scale GAN training for high fidelity natural image synthesis,” in Proc. Int. Conf. Learn. Represent., 2018.

[32] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980.

[33] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in PyTorch,” in Proc. 31st Conf. Neural Inf. Process. Syst., 2017.

[34] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and improving the image quality of StyleGAN,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 8110–8119.

[35] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, “Training generative adversarial networks with limited data,” in Advances in Neural Information Processing Systems, vol. 33, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds. Red Hook, NY, USA: Curran Associates, 2020, pp. 12104–12114.

[36] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2019, pp. 4401–4410.

[37] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2009, pp. 248–255.

[38] A. Borji, “Pros and cons of GAN evaluation measures,” Comput. Vis. Image Understand., vol. 179, pp. 41–65, Feb. 2019.

[39] T. Nguyen, T. Le, H. Vu, and D. Phung, “Dual discriminator generative adversarial nets,” in Proc. Adv. Neural Inf. Process. Syst., vol. 30, 2017.

[40] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” J. Mach. Learn. Res., vol. 15, pp. 1929–1958, Jun. 2014.

[41] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “Mixup: Beyond empirical risk minimization,” 2017, arXiv:1710.09412.

[42] S. Liu, O. Bousquet, and K. Chaudhuri, “Approximation and convergence properties of generative adversarial learning,” in Proc. 31st Int. Conf. Neural Inf. Process. Syst., 2017, pp. 5551–5559.