# Metric-based Generative Adversarial Network

Guoxian Dai
NYU Multimedia and Visual Computing Lab
Dept. of ECE, NYU Abu Dhabi, UAE
Dept. of CSE, NYU Tandon School of Engineering, USA

Jin Xie
NYU Multimedia and Visual Computing Lab
Dept. of ECE, NYU Abu Dhabi, UAE
Dept. of CSE, NYU Tandon School of Engineering, USA

Yi Fang*
NYU Multimedia and Visual Computing Lab
Dept. of ECE, NYU Abu Dhabi, UAE
Dept. of ECE, NYU Tandon School of Engineering, USA

## ABSTRACT

Existing methods of generative adversarial network (GAN) use different criteria to distinguish between real and fake samples, such as probability [9], energy [44] or other losses [30]. In this paper, by employing the merits of deep metric learning, we propose a novel metric-based generative adversarial network (MBGAN), which uses the distance-criteria to distinguish between real and fake samples. Specifically, the discriminator of MBGAN adopts a triplet structure and learns a deep nonlinear transformation, which maps input samples into a new feature space. In the transformed space, the distance between real samples is minimized, while the distance between real sample and fake sample is maximized. Similar to the adversarial procedure of existing GANs, a generator is trained to produce synthesized examples, which are close to real examples, while a discriminator is trained to maximize the distance between real and fake samples to a large margin. Meanwhile, instead of using a fixed margin, we adopt a data-dependent margin [30], so that the generator could focus on improving the synthesized samples with poor quality, instead of wasting energy on well-produce samples. Our proposed method is verified on various benchmarks, such as CIFAR-10, SVHN and CelebA, and generates high-quality samples.

## CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; *Unsupervised learning*; Neural networks; Adversarial learning;

## KEYWORDS

deep metric learning, generative adversarial network, data-dependent margin

## 1 INTRODUCTION

Deep learning has shown dominant superiority over various computer vision tasks, such as image classification [12, 35, 38], object detection [8, 33]. However, most of the existing deep learning methods are supervised, which heavily rely on large amounts of labeled data. And it is quite time-consuming and costs too much to achieve such massive amounts of labeled data. Recently, the unsupervised models, especially, generative adversarial network (GAN) [9] and variational autoencoder [17], have attracted more and more attentions from researchers. A traditional GAN generally consists of two components, a generator and a discriminator. The generator is trained to generate high-quality synthesized data from random noise, so that those synthesized data could fool the discriminator. The discriminator is trained to distinguish between real data and synthesized data. The two components keep competing each other and reach an equilibrium until the discriminator could not distinguish between real data and synthesized data.

The idea of GAN [9] is straightforward and nice, however, the training process of GAN is quite tricky, which is very vulnerable to collapse. Generally, it would be much easier to distinguish between real and synthesized data than generating high quality data to fool the discriminator. When the discriminator is so strong and the generator is so weak, the discriminator could easily distinguish between real data and synthesized data. Under such condition, the gradient from the discriminator would be almost 0, which could hardly help improve the generator, and the training process is collapsed. Various types of GAN are proposed to improve the performance, such as DCGAN [31], EBGAN [44], WGAN [1] and LSGAN [30]. DCGAN [31] proposed a series of architectural guidelines to construct stable deep convolutional GAN, such as batch normalization and leaky relu. Instead of outputting probabilities, the discriminator of EBGAN [44] adopted an autoencoder as an energy function to distinguish between real and fake data. WGAN [1] adopted wasserstein metric to improve the stability of GAN, while LSGAN [30] proposed a loss-sensitive model so that the network could focus on improving the synthesized data with poor quality.

In this work, we propose a novel metric-based generative adversarial network (MBGAN). Instead of using probability [9], energy [44] or other losses [30] to distinguish between real data and fake data, we view the discriminator as a deep nonlinear transformation, which maps input samples into a new feature space. In the transformed space, the distance between real samples is minimized, while the distance between real and fake samples is maximized to a large margin. In addition, instead of being fixed, the marginal distance between real data and fake data is adaptive to the quality of synthesized data, so that the generator could focus on improving the poor produced samples. The adversarial training procedure is similar to existing GANs, a generator is trained to produce samples close to the real data in the transformed space, while the discriminator is trained to maximize the distance between real data and fake data. Compared to existing GANs, our proposed MBGAN is more straightforward, jointly learning with real samples and fake samples together to guide the adversarial training process. For the

*Corresponding author. Email: yfang@nyu.edu


*MM '17, October 23–27, 2017, Mountain View, CA, USA*

ACM ISBN 978-1-4503-4906-2/17/10…$15.00
https://doi.org/10.1145/3123266.3123334

discriminator, we adopt triplet-wise examples to distinguish between real data and fake data; for the generator, we adopt pairwise examples to minimize the distance between real and fake examples.

The main contribution of this paper is to use the idea of deep metric learning to train GAN. Instead of outputting probability or energy, the discriminator of our proposed MBGAN simply outputs a feature vector as the representation of input examples in the transformed space. The distance of the representations for the real examples should be close to each other, while being away from those fake examples. The existing methods of GAN treat real data and fake data individually, either assigning different probabilities [9], different energies [44], or different losses [30]. However, our proposed method jointly considers real and fake data with triplet-wise and pairwise examples for training discriminator and generator respectively. Compared to traditional methods, our proposed MBGAN is more straightforward to directly use the real data to help improve the data with poor quality produced by the generator, meanwhile help the discriminator to distinguish between real data and fake data.

The rest of the paper is organized as follows. In Section 2, we introduce the related works. In Section 3, we present our proposed method, metric-based generative adversarial network. In Section 4, we verify the proposed MBGAN on various benchmarks. In Section 5, we conclude our paper.

## 2 RELATED WORK

The related works are introduced from two aspects, one is about generative adversarial network and the other is about deep metric learning. Next we will discuss the representative works for the above two aspects.

*Generative adversarial network.* Generative adversarial network was first proposed by Goodfellow *et al.* [9], which simultaneously trains a generator and a discriminator via an adversarial procedure. The idea of GAN is actually a minmax problem in game theory. Through the adversarial competition, the generator could produce contrast data from random noise, which follows similar distribution as real data. The idea of GAN is straightforward, however, the training procedure is not stable, and vulnerable to collapse. Radford *et al.* [31] extended the idea of GAN with deep convolutional neural network by employing a set of architectural guidelines on the structure of current CNN model, such as replacing pooling with strided-convolution, replacing ReLU with leaky ReLU. Denton *et al.* [7] stacked multiple-stage GANs in a laplacian pyramid framework to generate high quality images. Im *et al.* [16] proposed generative recurrent adversarial network, in which the generator consists of a recurrent loop to improve the quality of produced samples. Reed *et al.* [32] proposed a new GAN model, which could provide more detailed control over the synthesized images, such as the content and its location. In addition, Mathieu *et al.* [25] employed the adversarial training for video prediction. Except for synthesizing 2D images, Wu *et al.* [40] apply GAN with 3D convolution to synthesize high-quality 3D objects.

Traditional GANs use the probability criteria to distinguish between real samples and fake samples, which is vulnerable to collapse. Therefore, lots of other GAN models are proposed to increase its stabilities. Zhao *et al.* [44] proposed an energy-based GAN, which views the discriminator as an energy function, specifically, the reconstruction error of autoencoder. Arjovsky *et al.* [1] proposed a Wasserstein-GAN through minimizing the approximated earth mover distance, which shows more stable behavior compared to the traditional GAN. In addition, Qi *et al.* [30] proposed a loss-sensitive GAN, which allows the generator to focus on improving samples with poor quality. Through combining variational autoencoder [17] and GAN, Larsen *et al.* [20] proposed a VAE-GAN, which simultaneously learns to encode and decoder generator/discriminator. Similarly, Warde-Farley *et al.* [39] adopted denoising autoencoder to improve the performance of GAN. Nowozin *et al.* [27] demonstrated that any $f$-divergence can be used to train GAN, moreover [9] is just a special case. In addition, Chen *et al.* [3] proposed an InfoGAN, which aims to maximize the mutual information between latent variables and observations in addition to adversarial loss.

*Deep metric learning.* In spired by the great success of deep learning [12, 19, 35, 38], deep metric learning with the siamese structure [2] was proposed recently. Compared to traditional metric learning [5, 6, 10] with a linear transformation, the deep metric learning could learn a more powerful deep nonlinear transformation to minimize the distance of positive pairs and maximize the distance of negative pairs. Deep metric learning has been applied to various computer vision tasks, such as face verification [4, 14], visual tracking [15], image retrieval [22], person re-identification [41] and dimensionality reduction [11]. Different from the above methods with randomly selecting training pairs, Song *et al.* [28] considered all the possible pairs in a minibatch for metric learning. Apart from the siamese architecture, Hoffer *et al.* [13] proposed to use a triplet structure with triplet-wise examples, one anchor, one positive and one negative examples. Similarly, Sohn *et al.* [36] adopted a multi-class N-pair loss to enable faster convergence and achieve better performance. In addition, the contrastive loss is often combined with the classification loss to jointly train the network [23, 29, 37, 42] for further improving performance. Parkhi *et al.* [29] proposed a deep face model, which adopted softmax loss for face classification and triplet loss for face embedding. Similarly, Zhang *et al.* [43] jointly combined the classification loss and contrastive loss for learning fine-grained features.

## 3 PROPOSED METHOD

In this paper, we propose a novel metric-based generative adversarial network as shown Fig. 1. Instead of outputting probability [9], energy [44] or other losses [1, 30], we view the discriminator as a deep nonlinear transformation, which maps input samples into a new feature space. In the transformed space, the Euclidean distance between real samples should be encouraged as small as possible, meanwhile away from fake examples. Specifically, we adopt a triplet structure for feature embedding. The adversarial procedure is that the discriminator is trained to minimize the distance between real samples and maximize the distance between real and fake samples to a data-dependent margin, while the generator is trained to minimize the distance between real samples and fake samples.

Let $X = \{x_1, x_2, \cdots, x_N\}$ denote the set of training samples, $p_{data}$ denote the distribution of real samples, while $z$ is sampled from uniform distribution $U(0, 1)$. The generator is trained to map input noise $z$ into a sample $G(z)$. The discriminator learns a deep

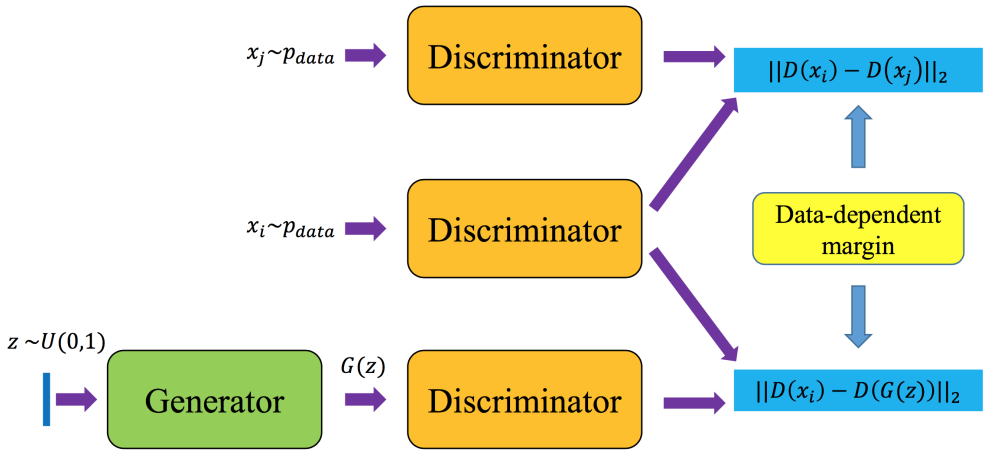



**Figure 1: Detailed framework of our proposed method. $z$ is sampled from uniform distribution $U(0,1)$. $G, D$ denote the transfer functions for the generator and discriminator respectively, which have the similar structures as DCGAN [31]. We view the discriminator as a deep nonlinear transformation, mapping input samples into a new feature space. In the transformed space, we use the Euclidean distances between the representations of different samples to distinguish real and fake samples. The adversarial procedure is that the discriminator is trained to minimize the pairwise distance among real samples, and maximize the pairwise distance between real data and fake data to a data-dependent margin. Meawhile the generator is trained to minimize the Euclidean distance between real data and fake data.**

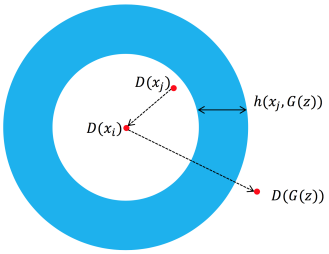



**Figure 2: Illustration of the proposed MBGAN. In the transformed space, the distance between real samples, $D(x_i)$ and $D(x_j)$ is pushed to be close to each other, while the distance between real sample $D(x_i)$ and fake sample $D(G(z))$ is push away from each other to a data-dependent margin.**

nonlinear transformation $D : x \rightarrow D(x)$, $D(x) \in \mathbb{R}^K$, mapping input samples into a new feature space, where $K$ is dimensionality of the new feature space.

### 3.1 Objective function

Our proposed method views the discriminator as a deep nonlinear transformation, as illustrated in Fig. 2. The adversarial procedure is actually adversarial distance-metric learning. The generator is trained to produce samples that are close to real samples in the new feature space, while the discriminator is trained to maximize the distance between real data and fake data to a data-dependent margin. To this end, the proposed generator loss $\mathcal{L}_G$ and discriminator loss $\mathcal{L}_D$ are defined as follows,

$$\mathcal{L}_D(x_i, x_j, z) = \max\{0, h(z, x_j) + d(x_i, x_j) - d(x_i, G(z))\} \quad (1)$$

$$\mathcal{L}_G(x_j, z) = d(x_j, G(z)) \quad (2)$$

where $d$ denotes the pairwise distance of samples in the transformed space, as shown in Eq. 3, $h$ denotes the data-dependent margin,

$$\begin{aligned} d(x_i, x_j) &= \|D(x_i) - D(x_j)\|_2 \\ d(x_i, G(z)) &= \|D(x_i) - D(G(z))\|_2 \\ d(x_j, G(z)) &= \|D(x_j) - D(G(z))\|_2 \end{aligned} \quad (3)$$

Instead of using a fixed margin $h$ for $\mathcal{L}_D$ in Eq. 1, we adopt a data-dependent margin [30]. The data-dependent margin could allow the generator to focus on improving poor-produced samples, instead of wasting effects on well-produced samples. Specifically, we use the $L_1$ distance between pairwise input samples, pixel-wise difference, which is balanced by weight parameter $\alpha$, shown as follows:

$$h(x_j, G(z)) = \alpha \|x_j - G(z)\|_1 \quad (4)$$

Overall, the proposed method could be optimized with back-propagation, by alternatively updating generator and discriminator, as shown in Algorithm 1.

**Algorithm 1** Training algorithm for MBGAN

**Input:** Training set $X = \{x_1, x_2, \cdots\}$; weight parameter $\alpha$.
**for** *iteration* = 0 to $M$ **do**
  Sample a minibatch $\{(x_i, x_j, z) | x_i \in X, x_j \in X, z \sim U(0,1)\}$
  Update discriminator loss $\mathcal{L}_D$ by descending the gradient of Eq. 1
  **for** *step* = 0 to $N$ **do**
    Sample a minibatch from $U(0,1)$.
    Update generator loss $\mathcal{L}_G$ by descending the gradient of Eq. 2
  **end for**
**end for**

### 3.2 Comparison with existing GANs

In this section, we will give a brief comparison between our proposed MBGAN with existing GANs, including GAN [9], EBGAN [44] and LSGAN [30]. The main difference lies with discriminator which adopts different criteria to distinguish between real samples and fake samples. The traditional GAN [9] adopts the probability criteria, thus the discriminator is trained to output high probability for real samples and low probability for fake data. The EBGAN [44] adopted a energy criteria, which views the discriminator as an energy function, and the discriminator is trained to output low energy for real data, and high energy for fake samples. The LSGAN [30] adopted a loss criteria, and arbitrary loss function could be applied, and the discriminator is trained to produce low loss for real data, and high loss for fake data. In addition, the margin between losses of real samples and fake samples is data-dependent, allowing the generator to focus on improving poor-produced data.

Different from the above methods, our proposed method adopts a distance-metric criteria to distinguish between real samples and fake samples. And we view the discriminator as a deep nonlinear transformation. In the transformed space, the representations of real samples should be close to each other, meanwhile away from fake data. The discriminator is trained to minimize the distance between real samples and maximize the distance between real data and fake data to data-dependent margin, similar as [30], while the generator is trained to minimize the distance between real data and fake data.

### 3.3 Conditional MBGAN

Our proposed MBGAN could be easily extended to conditional model by inputting extra information $y$. The extra information $y$ could be the class label, or text description, *etc.* And the generator is trained to produce desired images based on the prior information $y$. For our experiments, $y$ is the class label, which is a one-hot vector, along with the input noise vector $z$. The updated generator loss $\mathcal{L}_G$ and discriminator loss $\mathcal{L}_D$ are defined as follows:

$$\begin{aligned}\mathcal{L}_D(x_i, x_j, z|y) = \max\{0, h(z, x_j|y) + d(x_i, x_j|y) \\ - d(x_i, G(z)|y)\}\end{aligned} \tag{5}$$

$$\mathcal{L}_G(x_j, z|y) = d(x_j, G(z)|y) \tag{6}$$

The conditional model could provide more control of the produced samples for GAN.

## 4 EXPERIMENTAL RESULTS

Our proposed MBGAN is verified on various datasets, such as CIFAR-10 [18], street view house number (SVHN) [26] and CelebA [24]. And our proposed MBGAN could generate samples with high quality, compared with other GAN models. Except for visual comparisons, we also provide quantitative comparison by extracting features from discriminator and applying them with a supervised image classification task.

### 4.1 Implementation details

In this subsection, we briefly introduce the implementation details. We adopt similar structure as DCGAN [31]. The main difference is that the discriminator of MBGAN outputs a high-dimensional feature vector, while the discriminator of DCGAN produces the probability.

**Table 1: Structure of generator.**

| Input 100-D random noise |
|---|
| 5c2s512o UpConv. BN LeakyReLU |
| 5c2s256o UpConv. BN LeakyReLU |
| 5c2s128o UpConv. BN LeakyReLU |
| 5c2s64o UpConv. BN LeakyReLU |
| 5c2s3o UpConv. BN LeakyReLU |
| Elementwise Tanh |
| Output $64 \times 64 \times 64 \times 3$ |

**Table 2: Structure of discriminator.**

| Input $64 \times 64 \times 64 \times 3$ |
|---|
| 5c2s64o Conv. BN LeakyReLU |
| 5c2s128o Conv. BN LeakyReLU |
| 5c2s256o Conv. BN LeakyReLU |
| 5c2s512o Conv. BN LeakyReLU |
| 500o FC. |
| Output 500-D feature vector |

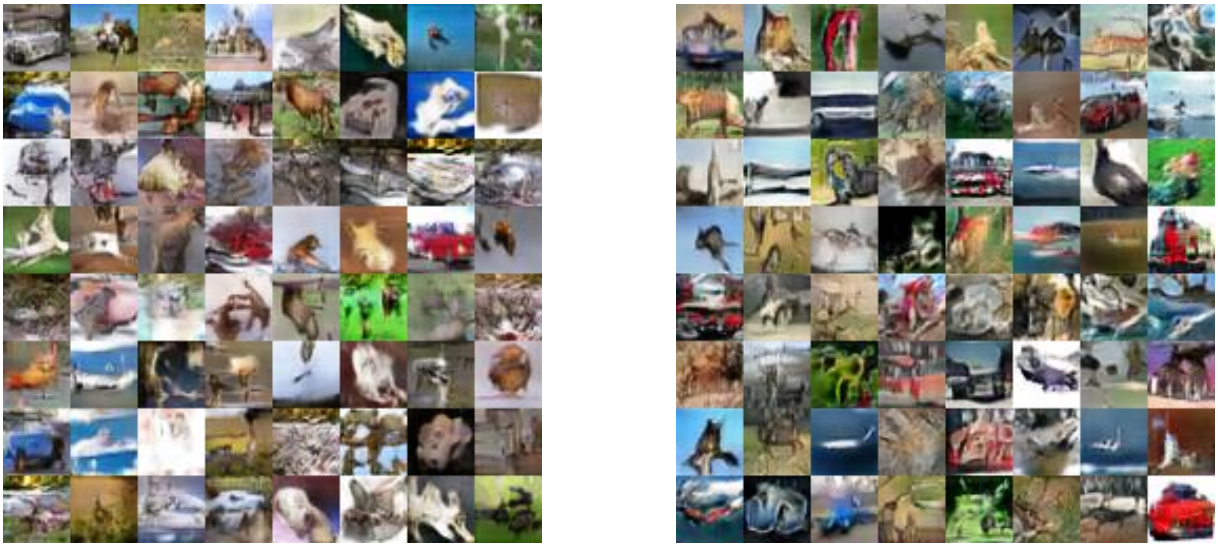
Tables. 1 and 2 show the detail structures of MBGAN for training on CelebA dataset, where BN stands for batch normalization, UpConv. for fractionally-strided convolution, FC for fully connected, and "5c2s512o" for $5 \times 5$ kernel with stride 2 and 512 outputs. All the faces from CelebA are cropped and resized to $64 \times 64$. In addition, the learning rate is set to 0.0002, $\beta_1$ for adam optimizer is set to 0.5 and batch size is set to 128.

As for CIFAR-10 and SVHN, the images are $32 \times 32$. We just make a slight change of structure in Tables. 1 and 2. The stride was changed from 2 to 1 for both the last convolution layer in the generator and first convolution layer in discriminator.

### 4.2 CIFAR-10

CIFAR-10 [18] contains 60000 $32 \times 32$ images, which are divided into 10 classes. For each class, there are 6000 images, 5000 for training and 1000 for testing. All the training data are used to train the proposed MBGAN.

Fig. 3 shows the generated images from both DCGAN and MBGAN on CIFAR-10 dataset. As we can see from Fig. 3, there are no visual differences between images generated from both DCGAN and MBGAN. Except for visual comparison, we also conduct quantitative comparison between DCGAN and MBGAN to evaluate the deep learned features of MBGAN. Specifically, we follow the same experimental setting in DCGAN [31], the activations from all the convolution layers of the discriminator are extracted. The extracted features are passed through an max-pooling operator to form a $4 \times 4$ grid, and they are concatenated to form one high-dimensional vector, with size of 18432. Finally, we train a regularized L2-SVM with the extracted high-dimensional representation. The performance comparison is listed in Table. 3. Our proposed MBGAN could slightly outperform DCGAN with the gain of 0.01.



(a) DCGAN (b) MBGAN

**Figure 3: Images generated by DCGAN and MBGAN on CIFAR-10 dataset.**

**Table 3: Performance comparison between DCGAN and MBGAN on CIFAR-10 dataset.**

| Methods | Accuracy |
|---|---|
| DCGAN [31] | 0.758 |
| MBGAN | 0.768 |

It is noted that the reported classification accuracy of DCGAN on CIFAR-10 in [31] is 0.828, which is trained on ImageNet [34] while tested on CIFAR-10 dataset. The ImageNet [34] contains more than 1M images with 1K classes, which is much larger than CIFAR10 with 60K images of 10 classes. Thus the model trained on ImageNet tends to have much more powerful generalization capability than the model trained on CIFAR-10. Therefore, as a fair comparison, we retrain DCGAN on CIFAR-10, and extract features from the discriminator of DCGAN. The extracted features are used to train a linear L2-SVM, and the reimplemented classification accuracy on CIFAR-10 is only 0.758.

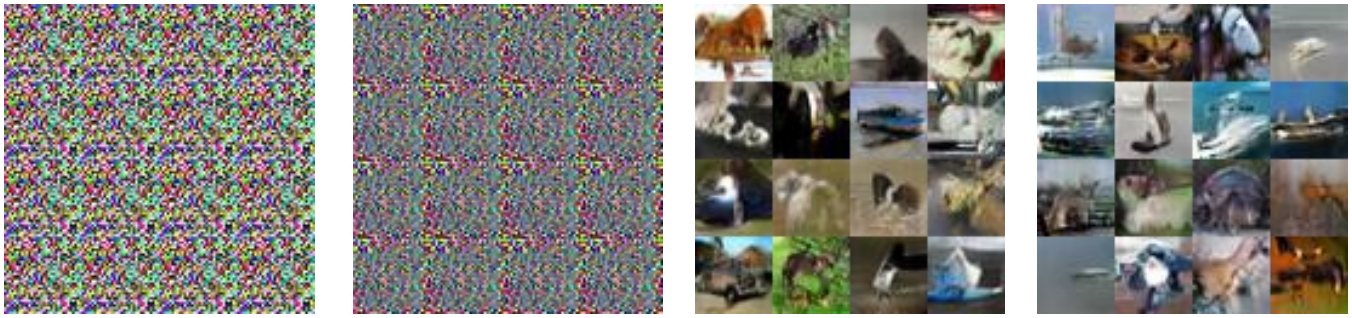
Compared to traditional DCGAN [31], our proposed MBGAN has two additional parameters $\alpha$ and $K$, where $\alpha$ is used to control the magnitude of the data-dependent margin $h$, while $K$ is dimensionality of the output feature space. To fully explore our proposed MBGAN, we verify the effects of both $\alpha$ and $K$ to the training process. For verifying the effects of $\alpha$, we fix $K = 500$ and choose 4 different $\alpha$ for comparison, namely 0.2, 2, 20, 200. The generated images are listed as shown in Fig. 4. As we can see in Fig. 4, when $\alpha$ is small, namely 0.2, 2, the model collapses and the generated images are just noise. When $\alpha$ is larger, namely 20, 200, the model is much more stable, and generates nice images. The experimental results are reasonable since a larger margin makes it easier for the discriminator to distinguish between real and fake samples, and vice versa. For verifying the effects of $K$, we fix $\alpha = 20$, and choose four different $K$, namely 5, 50, 500, 5000. The generated images are listed as follows in Fig. 5. As we can see in Fig. 5, when the output feature dimensionality $K$ is small, namely 5, 50, the model collapses and the generated images are just noise. When $K$ is larger, namely 500, 5000, the model is much more stable, and generates nice images. The experimental results are reasonable since features with larger dimensions are easier for the discriminator to distinguish between real and fake samples, and vice versa.

In addition, we also analyze the magnitude ($l_2$ norm) of the gradients for the generator during training. Generally, the traditional GAN is vulnerable to collapse, and the gradients of the generator tend to vanish, which is the main problem for GAN training. Fig. 6 shows the magnitude of the gradients for the generator of MBGAN over interations. As we can see in Fig. 6, the magnitude of the generator's gradients steadily remains above 0 with a large margin. During training, our proposed MBGAN quickly reaches equilibrium after a few epochs. Even if the model reaches an equilibrium, it can still provide sufficient gradient to continuously update the generator of MBGAN.

## 4.3 Street view house number dataset

Street view house number (SVHN) dataset [26] is obtained from Google street view images. Similar to MNIST [21], there are 10 classes in SVHN dataset, including digits '0-9'. There are 73257 digits for training, 26032 digits for testing and 531131 digits as extra training data for future improving performance. All the images are cropped and resized to a fixed resolution of $32 \times 32$ pixels. All the training data are used to train the proposed MBGAN.

Fig. 7 shows the images generated by both DCGAN and MBGAN on SVHN dataset. As we can see from Fig. 7, there is no distinct visual difference between the images generated by both types of GANs. To quantitatively measure the performance of the proposed MBGAN, we use the same experiment setup as CIFAR-10. We first



**(a)** $\alpha = 0.2$ **(b)** $\alpha = 2$ **(c)** $\alpha = 20$ **(d)** $\alpha = 200$

**Figure 4: The effects of different margins to the images generated by MBGAN on CIFAR-10 dataset.**

**(a)** $K = 5$ **(b)** $K = 50$ **(c)** $K = 500$ **(d)** $K = 5000$

**Figure 5: The effects of different dimensions for output features to the images generated by MBGAN on CIFAR-10 dataset.**

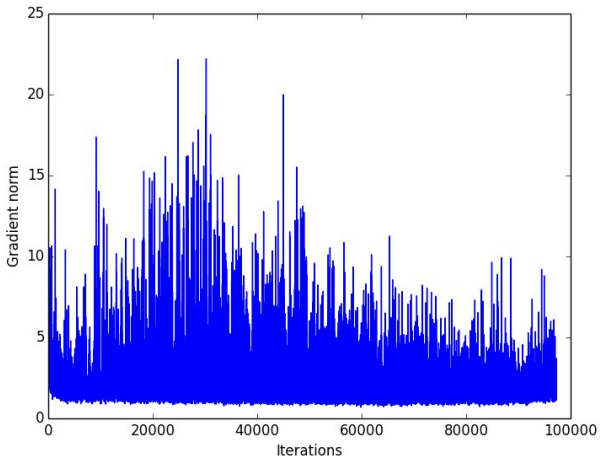



**Figure 6: The gradient norm of MBGAN's generator over iterations.**

extract deep features from the convolutional layers of discriminator and then train a linear L2-SVM for classification on SVHN dataset. Table. 4 shows the performance comparison between DCGAN and MBGAN on SVHN dataset. As we can see from Table. 4, the proposed MBGAN could achieve comparable performance with DCGAN.

**Table 4: Performance comparison between DCGAN and MBGAN on SVHN dataset.**

| Methods | Accuracy |
|---|---|
| DCGAN [31] | 0.890 |
| MBGAN | 0.877 |

## 4.4 CelebA dataset

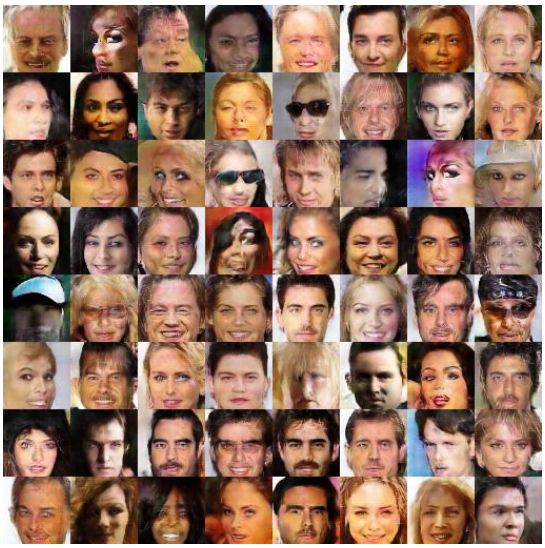
CelebFaces Attribute Dataset (CelebA) [24] is a large scale face attributes dataset. The CelebA contains 202599 number of face images, which are from 10177 celebrity identities. For each image, there are 5 landmark locations and 40 binary attributes. Fig. 8 shows the images generated by both DCGAN and MBGAN on CelebA dataset. There is no significant difference between images generated from both models.

To demonstrate the generalization capability of the proposed MBGAN, instead of simply memorizing the training images, we interpolate the random noise $z$ and map it to the synthesized image as shown in Fig. 9. The images of both the leftmost and rightmost columns are generated from random noises $z_l$ and $z_r$, while the images between them are generated from the linear interpolations of their corresponding noise vectors, $z_m = \lambda z_l + (1 - \lambda) z_r$, where

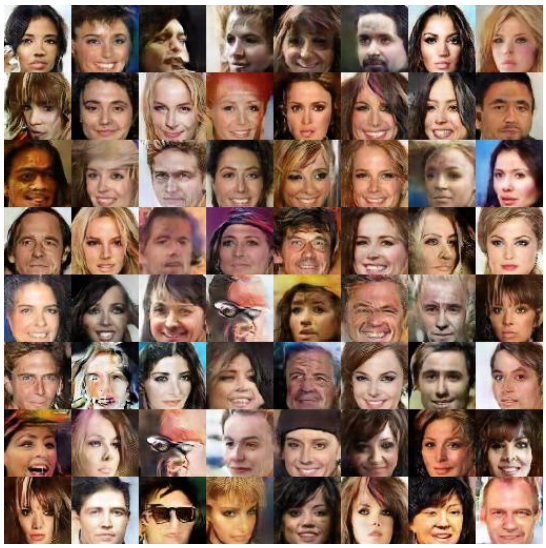
(a) DCGAN

(b) MBGAN

**Figure 7: Images generated by DCGAN and MBGAN on SVHN dataset.**



(a) DCGAN



(b) MBGAN

**Figure 8: Images generated by DCGAN and MBGAN on CelebA dataset.**

$\lambda \in [0, 1]$. As we can see from Fig. 9, the generated images slowly transition between leftmost and rightmost column. For example, on the first row, the color of the hair slowly transitions from black to gold, in addition, the hair style and face color also change smoothly. On the third row, a woman's face with long hair and open mouth slowly transitions into a man's face with short hair and closed mouth. On the last row, a women without sunglass slowly transitions into wearing sunglass. During transition, the region around eyes become darker and darker smoothly. The interpolated images demonstrate the smooth continuity of the proposed MBGAN.

## 5 CONCLUSION

In this work, by employing deep metric learning, we proposed a novel metric-based generative adversarial network. Different from existing methods using probability criteria, energy criteria or other loss criteria, we adopted distance-criteria for the discriminator to distinguish between real samples and fake samples. Specifically, the discriminator adopts a triplet structure with triplet-wise input examples and learns a deep nonlinear transformation, which maps input samples from the original space into a new feature space. In the transformed space, a generator is trained to minimize the

**Figure 9: The images generated by the linear interpolation.**

distance between real sample and fake sample, while a discriminator is trained to maximize the distance between real sample and fake sample to a data-dependent margin. The data-dependent margin could allow the generator to focus on improving images with poor quality, instead of wasting energy on images with high quality. Finally, our proposed method is verified on various datasets and generates high-quality images.

## 6 ACKNOWLEDGMENTS

This material is partly based upon work supported by New York University Abu Dhabi institute (AD131 and REF131).

## REFERENCES

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein gan. *arXiv preprint arXiv:1701.07875* (2017).
[2] Jane Bromley, James W. Bentz, Léon Bottou, Isabelle Guyon, Yann LeCun, Cliff Moore, Eduard Säckinger, and Roopak Shah. 1993. Signature Verification Using A "Siamese" Time Delay Neural Network. *IJPRAI* 7, 4 (1993), 669–688.
[3] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*. 2172–2180.
[4] Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, Vol. 1. IEEE, 539–546.
[5] Zhen Cui, Wen Li, Dong Xu, Shiguang Shan, and Xilin Chen. 2013. Fusing robust face region descriptors via multiple metric learning for face recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3554–3561.
[6] Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon. 2007. Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning*. ACM, 209–216.
[7] Emily L Denton, Soumith Chintala, Rob Fergus, et al. 2015. Deep Generative Image Models using a? Laplacian Pyramid of Adversarial Networks. In *Advances in neural information processing systems*. 1486–1494.
[8] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 580–587.
[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.
[10] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. 2009. Is that you? Metric learning approaches for face identification. In *Computer Vision, 2009 IEEE 12th international conference on*. IEEE, 498–505.
[11] Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, Vol. 2. IEEE, 1735–1742.
[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385* (2015).
[13] Elad Hoffer and Nir Ailon. 2015. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*. Springer, 84–92.
[14] Junlin Hu, Jiwen Lu, and Yap-Peng Tan. 2014. Discriminative deep metric learning for face verification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1875–1882.
[15] Junlin Hu, Jiwen Lu, and Yap-Peng Tan. 2016. Deep metric learning for visual tracking. *IEEE Transactions on Circuits and Systems for Video Technology* 26, 11 (2016), 2056–2068.
[16] Daniel Jiwoong Im, Chris Dongjoo Kim, Hui Jiang, and Roland Memisevic. 2016. Generating images with recurrent adversarial networks. *arXiv preprint arXiv:1602.05110* (2016).
[17] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).
[18] Alex Krizhevsky and Geoffrey Hinton. 2009. Learning multiple layers of features from tiny images. (2009).
[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.
[20] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. 2015. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300* (2015).
[21] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.
[22] Zechao Li and Jinhui Tang. 2015. Weakly supervised deep metric learning for community-contributed image retrieval. *IEEE Transactions on Multimedia* 17, 11 (2015), 1989–1999.
[23] Jingtuo Liu, Yafeng Deng, Tao Bai, Zhengping Wei, and Chang Huang. 2015. Targeting ultimate accuracy: Face recognition via deep embedding. *arXiv preprint arXiv:1506.07310* (2015).
[24] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.
[25] Michael Mathieu, Camille Couprie, and Yann LeCun. 2015. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440* (2015).
[26] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. 2011. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, Vol. 2011. 5.

[27] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. 2016. f-GAN: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*. 271–279.
[28] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. 2016. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4004–4012.
[29] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. 2015. Deep Face Recognition.. In *BMVC*, Vol. 1. 6.
[30] Guo-Jun Qi. 2017. Loss-Sensitive Generative Adversarial Networks on Lipschitz Densities. *arXiv preprint arXiv:1701.06264* (2017).
[31] Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434* (2015).
[32] Scott E Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. 2016. Learning what and where to draw. In *Advances In Neural Information Processing Systems*. 217–225.
[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.
[34] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115, 3 (2015), 211–252. https://doi.org/10.1007/s11263-015-0816-y
[35] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
[36] Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*. 1849–1857.
[37] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. 2014. Deep learning face representation by joint identification-verification. In *Advances in neural information processing systems*. 1988–1996.
[38] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1–9.
[39] David Warde-Farley and Yoshua Bengio. 2017. Improving generative adversarial networks with denoising feature matching. (2017).
[40] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. 2016. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Advances in Neural Information Processing Systems*. 82–90.
[41] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. 2014. Deep metric learning for person re-identification. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*. IEEE, 34–39.
[42] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. 2014. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923* (2014).
[43] Xiaofan Zhang, Feng Zhou, Yuanqing Lin, and Shaoting Zhang. 2016. Embedding label structures for fine-grained feature representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1114–1123.
[44] Junbo Zhao, Michael Mathieu, and Yann LeCun. 2016. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126* (2016).