
Triangle Generative Adversarial Networks

Zhe Gan*, **Liqun Chen***, **Weiyao Wang**, **Yunchen Pu**, **Yizhe Zhang**, **Hao Liu**, **Lawrence Carin**
Duke University
{zg27, lc267, ww109, yp42, yizhe.zhang, hao.liu1, lcarin}@duke.edu

Abstract

A Triangle Generative Adversarial Network (Δ -GAN) is developed for semi-supervised cross-domain joint distribution matching, where the training data consists of samples from each domain, and supervision of domain correspondence is provided by only a few *paired* samples. Δ -GAN consists of four neural networks, two generators and two discriminators. The generators are designed to learn the two-way conditional distributions between the two domains, while the discriminators implicitly define a ternary discriminative function, which is trained to distinguish real data pairs and two kinds of fake data pairs. The generators and discriminators are trained together using adversarial learning. Under mild assumptions, in theory the joint distributions characterized by the two generators concentrate to the data distribution. In experiments, three different kinds of domain pairs are considered, image-label, image-image and image-attribute pairs. Experiments on semi-supervised image classification, image-to-image translation and attribute-based image generation demonstrate the superiority of the proposed approach.

1 Introduction

Generative adversarial networks (GANs) [1] have emerged as a powerful framework for learning generative models of arbitrarily complex data distributions. When trained on datasets of natural images, significant progress has been made on generating realistic and sharp-looking images [2, 3]. The original GAN formulation was designed to learn the data distribution in one domain. In practice, one may also be interested in matching two joint distributions. This is an important task, since mapping data samples from one domain to another has a wide range of applications. For instance, matching the joint distribution of image-text pairs allows simultaneous image captioning and text-conditional image generation [4], while image-to-image translation [5] is another challenging problem that requires matching the joint distribution of image-image pairs.

In this work, we are interested in designing a GAN framework to match joint distributions. If paired data are available, a simple approach to achieve this is to train a conditional GAN model [4, 6], from which a joint distribution is readily manifested and can be matched to the empirical joint distribution provided by the paired data. However, fully supervised data are often difficult to acquire. Several methods have been proposed to achieve unsupervised joint distribution matching without any paired data, including Adversarially Learned Inference (ALI) [7], Bidirectional GAN (BiGAN) [8], DiscoGAN [9], CycleGAN [10] and DualGAN [11] among many other related works. Though empirically achieving great success, in principle, there exist infinite many possible mapping functions that satisfy the requirement to map a sample from one domain to another. In order to alleviate this nonidentifiability issue, paired data are needed to provide proper supervision to inform the model the kind of joint distributions that are desired.

* Equal contribution.

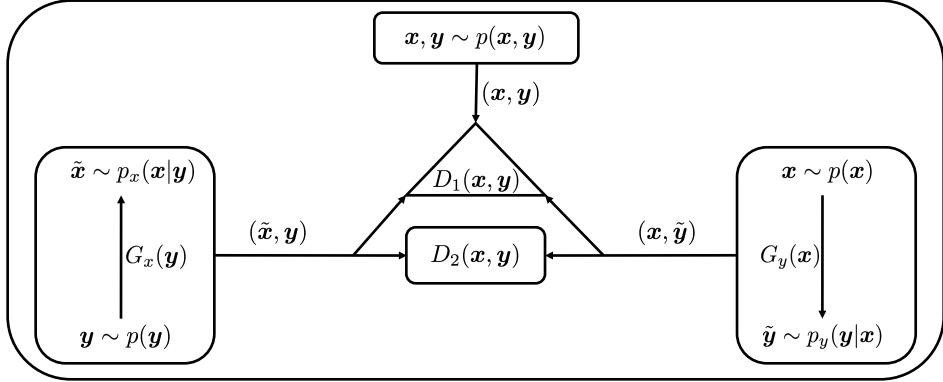


Figure 1: Illustration of the Triangle Generative Adversarial Network (Δ -GAN).

This motivates the proposed Triangle Generative Adversarial Network (Δ -GAN), a GAN framework that allows semi-supervised joint distribution matching, where the supervision of domain correspondence is provided by a few paired samples. Δ -GAN consists of two generators and two discriminators. The generators are designed to learn the bidirectional mappings between domains, while the discriminators are trained to distinguish real data pairs and two kinds of fake data pairs. Both the generators and discriminators are trained together via adversarial learning.

Δ -GAN bears close resemblance to Triple GAN [12], a recently proposed method that can also be utilized for semi-supervised joint distribution mapping. However, there exist several key differences that make our work unique. First, Δ -GAN uses two discriminators in total, which implicitly defines a ternary discriminative function, instead of a binary discriminator as used in Triple GAN. Second, Δ -GAN can be considered as a combination of conditional GAN and ALI, while Triple GAN consists of two conditional GANs. Third, the distributions characterized by the two generators in both Δ -GAN and Triple GAN concentrate to the data distribution in theory. However, when the discriminator is optimal, the objective of Δ -GAN becomes the Jensen-Shannon divergence (JSD) among three distributions, which is symmetric; the objective of Triple GAN consists of a JSD term plus a Kullback-Leibler (KL) divergence term. The asymmetry of the KL term makes Triple GAN more prone to generating fake-looking samples [13]. Lastly, the calculation of the additional KL term in Triple GAN is equivalent to calculating a supervised loss, which requires the explicit density form of the conditional distributions, which is generally not available. This makes the application of Triple GAN limited. On the other hand, Δ -GAN is a fully adversarial approach that does not require that the conditional densities can be computed; Δ -GAN only require that the conditional densities can be sampled from in a way that allows gradient backpropagation.

Δ -GAN is a general framework, and can be used to match any joint distributions. In experiments, in order to demonstrate the versatility of the proposed model, we consider three domain pairs: image-label, image-image and image-attribute pairs, and use them for semi-supervised classification, image-to-image translation and attribute-based image editing, respectively. In order to demonstrate the scalability of the model to large and complex datasets, we also present attribute-conditional image generation on the COCO dataset [14].

2 Model

2.1 Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) [1] consist of a generator G and a discriminator D that compete in a two-player minimax game, where the generator is learned to map samples from an arbitrary latent distribution to data, while the discriminator tries to distinguish between real and generated samples. The goal of the generator is to “fool” the discriminator by producing samples that are as close to real data as possible. Specifically, D and G are learned as

$$\min_G \max_D V(D, G) = \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x})} [\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim p_z(\mathbf{z})} [\log(1 - D(G(\mathbf{z})))] \quad (1)$$

where $p(\mathbf{x})$ is the true data distribution, and $p_z(\mathbf{z})$ is usually defined to be a simple distribution, such as the standard normal distribution. The generator G implicitly defines a probability distribution

$p_g(\mathbf{x})$ as the distribution of the samples $G(\mathbf{z})$ obtained when $\mathbf{z} \sim p_z(\mathbf{z})$. For any fixed generator G , the optimal discriminator is $D(\mathbf{x}) = \frac{p(\mathbf{x})}{p_g(\mathbf{x})+p(\mathbf{x})}$. When the discriminator is optimal, solving this adversarial game is equivalent to minimizing the Jensen-Shannon Divergence (JSD) between $p(\mathbf{x})$ and $p_g(\mathbf{x})$ [1]. The global equilibrium is achieved if and only if $p(\mathbf{x}) = p_g(\mathbf{x})$.

2.2 Triangle Generative Adversarial Networks (Δ -GANs)

We now extend GAN to Δ -GAN for joint distribution matching. We first consider Δ -GAN in the supervised setting, and then discuss semi-supervised learning in Section 2.4. Consider two related domains, with \mathbf{x} and \mathbf{y} being the data samples for each domain. We have fully-paired data samples that are characterized by the joint distribution $p(\mathbf{x}, \mathbf{y})$, which also implies that samples from both the marginal $p(\mathbf{x})$ and $p(\mathbf{y})$ can be easily obtained.

Δ -GAN consists of two generators: (i) a generator $G_x(\mathbf{y})$ that defines the conditional distribution $p_x(\mathbf{x}|\mathbf{y})$, and (ii) a generator $G_y(\mathbf{x})$ that characterizes the conditional distribution in the other direction $p_y(\mathbf{y}|\mathbf{x})$. $G_x(\mathbf{y})$ and $G_y(\mathbf{x})$ may also implicitly contain a random latent variable \mathbf{z} as input, i.e., $G_x(\mathbf{y}, \mathbf{z})$ and $G_y(\mathbf{x}, \mathbf{z})$. In the Δ -GAN game, after a sample \mathbf{x} is drawn from $p(\mathbf{x})$, the generator G_y produces a pseudo sample $\tilde{\mathbf{y}}$ following the conditional distribution $p_y(\mathbf{y}|\mathbf{x})$. Hence, the fake data pair $(\mathbf{x}, \tilde{\mathbf{y}})$ is a sample from the joint distribution $p_y(\mathbf{x}, \mathbf{y}) = p_y(\mathbf{y}|\mathbf{x})p(\mathbf{x})$. Similarly, a fake data pair $(\tilde{\mathbf{x}}, \mathbf{y})$ can be sampled from the generator G_x by first drawing \mathbf{y} from $p(\mathbf{y})$ and then drawing $\tilde{\mathbf{x}}$ from $p_x(\mathbf{x}|\mathbf{y})$; hence $(\tilde{\mathbf{x}}, \mathbf{y})$ is sampled from the joint distribution $p_x(\mathbf{x}, \mathbf{y}) = p_x(\mathbf{x}|\mathbf{y})p(\mathbf{y})$. As such, the generative process between $p_x(\mathbf{x}, \mathbf{y})$ and $p_y(\mathbf{x}, \mathbf{y})$ is reversed.

The objective of Δ -GAN is to match the three joint distributions: $p(\mathbf{x}, \mathbf{y})$, $p_x(\mathbf{x}, \mathbf{y})$ and $p_y(\mathbf{x}, \mathbf{y})$. If this is achieved, we are ensured that we have learned a bidirectional mapping $p_x(\mathbf{x}|\mathbf{y})$ and $p_y(\mathbf{y}|\mathbf{x})$ that guarantees the generated fake data pairs $(\tilde{\mathbf{x}}, \mathbf{y})$ and $(\mathbf{x}, \tilde{\mathbf{y}})$ are indistinguishable from the true data pairs (\mathbf{x}, \mathbf{y}) . In order to match the joint distributions, an adversarial game is played. Joint pairs are drawn from three distributions: $p(\mathbf{x}, \mathbf{y})$, $p_x(\mathbf{x}, \mathbf{y})$ or $p_y(\mathbf{x}, \mathbf{y})$, and two discriminator networks are learned to discriminate among the three, while the two conditional generator networks are trained to fool the discriminators.

The value function describing the game is given by

$$\begin{aligned} \min_{G_x, G_y} \max_{D_1, D_2} V(G_x, G_y, D_1, D_2) &= \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim p(\mathbf{x}, \mathbf{y})} [\log D_1(\mathbf{x}, \mathbf{y})] \\ &\quad + \mathbb{E}_{\mathbf{y} \sim p(\mathbf{y}), \tilde{\mathbf{x}} \sim p_x(\mathbf{x}|\mathbf{y})} \left[\log \left((1 - D_1(\tilde{\mathbf{x}}, \mathbf{y})) \cdot D_2(\tilde{\mathbf{x}}, \mathbf{y}) \right) \right] \\ &\quad + \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x}), \tilde{\mathbf{y}} \sim p_y(\mathbf{y}|\mathbf{x})} \left[\log \left((1 - D_1(\mathbf{x}, \tilde{\mathbf{y}})) \cdot (1 - D_2(\mathbf{x}, \tilde{\mathbf{y}})) \right) \right] \end{aligned} \quad (2)$$

The discriminator D_1 is used to distinguish whether a sample pair is from $p(\mathbf{x}, \mathbf{y})$ or not, if this sample pair is not from $p(\mathbf{x}, \mathbf{y})$, another discriminator D_2 is used to distinguish whether this sample pair is from $p_x(\mathbf{x}, \mathbf{y})$ or $p_y(\mathbf{x}, \mathbf{y})$. D_1 and D_2 work cooperatively, and the use of both implicitly defines a ternary discriminative function D that distinguish sample pairs in three ways. See Figure 1 for an illustration of the adversarial game and Appendix B for an algorithmic description of the training procedure.

2.3 Theoretical analysis

Δ -GAN shares many of the theoretical properties of GANs [1]. We first consider the optimal discriminators D_1 and D_2 for any given generator G_x and G_y . These optimal discriminators then allow reformulation of objective (2), which reduces to the Jensen-Shannon divergence among the joint distribution $p(\mathbf{x}, \mathbf{y})$, $p_x(\mathbf{x}, \mathbf{y})$ and $p_y(\mathbf{x}, \mathbf{y})$.

Proposition 1. *For any fixed generator G_x and G_y , the optimal discriminator D_1 and D_2 of the game defined by $V(G_x, G_y, D_1, D_2)$ is*

$$D_1^*(\mathbf{x}, \mathbf{y}) = \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x}, \mathbf{y}) + p_x(\mathbf{x}, \mathbf{y}) + p_y(\mathbf{x}, \mathbf{y})}, \quad D_2^*(\mathbf{x}, \mathbf{y}) = \frac{p_x(\mathbf{x}, \mathbf{y})}{p_x(\mathbf{x}, \mathbf{y}) + p_y(\mathbf{x}, \mathbf{y})}$$

Proof. The proof is a straightforward extension of the proof in [1]. See Appendix A for details. \square

Proposition 2. The equilibrium of $V(G_x, G_y, D_1, D_2)$ is achieved if and only if $p(\mathbf{x}, \mathbf{y}) = p_x(\mathbf{x}, \mathbf{y}) = p_y(\mathbf{x}, \mathbf{y})$ with $D_1^*(\mathbf{x}, \mathbf{y}) = \frac{1}{3}$ and $D_2^*(\mathbf{x}, \mathbf{y}) = \frac{1}{2}$, and the optimum value is $-3 \log 3$.

Proof. Given the optimal $D_1^*(\mathbf{x}, \mathbf{y})$ and $D_2^*(\mathbf{x}, \mathbf{y})$, the minimax game can be reformulated as:

$$C(G_x, G_y) = \max_{D_1, D_2} V(G_x, G_y, D_1, D_2) \quad (3)$$

$$= -3 \log 3 + 3 \cdot JSD(p(\mathbf{x}, \mathbf{y}), p_x(\mathbf{x}, \mathbf{y}), p_y(\mathbf{x}, \mathbf{y})) \geq -3 \log 3 \quad (4)$$

where JSD denotes the Jensen-Shannon divergence (JSD) among three distributions. See Appendix A for details. \square

Since $p(\mathbf{x}, \mathbf{y}) = p_x(\mathbf{x}, \mathbf{y}) = p_y(\mathbf{x}, \mathbf{y})$ can be achieved in theory, it can be readily seen that the learned conditional generators can reveal the true conditional distributions underlying the data, *i.e.*, $p_x(\mathbf{x}|\mathbf{y}) = p(\mathbf{x}|\mathbf{y})$ and $p_y(\mathbf{y}|\mathbf{x}) = p(\mathbf{y}|\mathbf{x})$.

2.4 Semi-supervised learning

In order to further understand Δ -GAN, we write (2) as

$$V = \underbrace{\mathbb{E}_{p(\mathbf{x}, \mathbf{y})}[\log D_1(\mathbf{x}, \mathbf{y})] + \mathbb{E}_{p_x(\tilde{\mathbf{x}}, \mathbf{y})}[\log(1 - D_1(\tilde{\mathbf{x}}, \mathbf{y}))] + \mathbb{E}_{p_y(\mathbf{x}, \tilde{\mathbf{y}})}[\log(1 - D_1(\mathbf{x}, \tilde{\mathbf{y}}))]}_{\text{conditional GAN}} \quad (5)$$

$$+ \underbrace{\mathbb{E}_{p_x(\tilde{\mathbf{x}}, \mathbf{y})}[\log D_2(\tilde{\mathbf{x}}, \mathbf{y})] + \mathbb{E}_{p_y(\mathbf{x}, \tilde{\mathbf{y}})}[\log(1 - D_2(\mathbf{x}, \tilde{\mathbf{y}}))]}_{\text{BiGAN/ALI}} \quad (6)$$

The objective of Δ -GAN is a combination of the objectives of conditional GAN and BiGAN. The BiGAN part matches two joint distributions: $p_x(\mathbf{x}, \mathbf{y})$ and $p_y(\mathbf{x}, \mathbf{y})$, while the conditional GAN part provides the supervision signal to notify the BiGAN part what joint distribution to match. Therefore, Δ -GAN provides a natural way to perform semi-supervised learning, since the conditional GAN part and the BiGAN part can be used to account for paired and unpaired data, respectively.

However, when doing semi-supervised learning, there is also one potential problem that we need to be cautious about. The theoretical analysis in Section 2.3 is based on the assumption that the dataset is fully supervised, *i.e.*, we have the ground-truth joint distribution $p(\mathbf{x}, \mathbf{y})$ and marginal distributions $p(\mathbf{x})$ and $p(\mathbf{y})$. In the semi-supervised setting, $p(\mathbf{x})$ and $p(\mathbf{y})$ are still available but $p(\mathbf{x}, \mathbf{y})$ is not. We can only obtain the joint distribution $p_l(\mathbf{x}, \mathbf{y})$ characterized by the few paired data samples. Hence, in the semi-supervised setting, $p_x(\mathbf{x}, \mathbf{y})$ and $p_y(\mathbf{x}, \mathbf{y})$ will try to concentrate to the empirical distribution $p_l(\mathbf{x}, \mathbf{y})$. We make the assumption that $p_l(\mathbf{x}, \mathbf{y}) \approx p(\mathbf{x}, \mathbf{y})$, *i.e.*, the paired data can roughly characterize the whole dataset. For example, in the semi-supervised classification problem, one usually strive to make sure that labels are equally distributed among the labeled dataset.

2.5 Relation to Triple GAN

Δ -GAN is closely related to Triple GAN [12]. Below we review Triple GAN and then discuss the main differences. The value function of Triple GAN is defined as follows:

$$V = \mathbb{E}_{p(\mathbf{x}, \mathbf{y})}[\log D(\mathbf{x}, \mathbf{y})] + (1 - \alpha)\mathbb{E}_{p_x(\tilde{\mathbf{x}}, \mathbf{y})}[\log(1 - D(\tilde{\mathbf{x}}, \mathbf{y}))] + \alpha\mathbb{E}_{p_y(\mathbf{x}, \tilde{\mathbf{y}})}[\log(1 - D(\mathbf{x}, \tilde{\mathbf{y}}))] + \mathbb{E}_{p(\mathbf{x}, \mathbf{y})}[-\log p_y(\mathbf{y}|\mathbf{x})] \quad (7)$$

where $\alpha \in (0, 1)$ is a constant that controls the relative importance of the two generators. Let Triple GAN-s denote a simplified Triple GAN model with only the first three terms. As can be seen, Triple GAN-s can be considered as a combination of two conditional GANs, with the importance of each conditional GAN weighted by α . It can be proven that Triple GAN-s achieves equilibrium if and only if $p(\mathbf{x}, \mathbf{y}) = (1 - \alpha)p_x(\mathbf{x}, \mathbf{y}) + \alpha p_y(\mathbf{x}, \mathbf{y})$, which is not desirable. To address this problem, in Triple GAN a standard supervised loss $\mathcal{R}_L = \mathbb{E}_{p(\mathbf{x}, \mathbf{y})}[-\log p_y(\mathbf{y}|\mathbf{x})]$ is added. As a result, when the discriminator is optimal, the cost function in Triple GAN becomes:

$$2JSD(p(\mathbf{x}, \mathbf{y}) || ((1 - \alpha)p_x(\mathbf{x}, \mathbf{y}) + \alpha p_y(\mathbf{x}, \mathbf{y}))) + KL(p(\mathbf{x}, \mathbf{y}) || p_y(\mathbf{x}, \mathbf{y})) + \text{const.} \quad (8)$$

This cost function has the good property that it has a unique minimum at $p(\mathbf{x}, \mathbf{y}) = p_x(\mathbf{x}, \mathbf{y}) = p_y(\mathbf{x}, \mathbf{y})$. However, the objective becomes asymmetrical. The second KL term pays low cost for generating fake-looking samples [13]. By contrast Δ -GAN directly optimizes the *symmetric* Jensen-Shannon divergence among three distributions. More importantly, the calculation of $\mathbb{E}_{p(\mathbf{x}, \mathbf{y})}[-\log p_y(\mathbf{y}|\mathbf{x})]$ in Triple GAN also implies that the explicit density form of $p_y(\mathbf{y}|\mathbf{x})$ is available, which is generally not true. On the other hand, Δ -GAN only requires that $p_y(\mathbf{y}|\mathbf{x})$ can be sampled from. For example, if $p_y(\mathbf{y}|\mathbf{x}) = \int \delta(\mathbf{y} - G_y(\mathbf{x}, \mathbf{z}))p(\mathbf{z})d\mathbf{z}$, and $\delta(\cdot)$ is the Dirac delta function, we can sample \mathbf{y} through sampling \mathbf{z} , however, the density function of $p_y(\mathbf{y}|\mathbf{x})$ is not explicitly available, therefore Δ -GAN can be implemented, while Triple GAN cannot.

2.6 Applications

Δ -GAN is a general framework that can be used for any joint distribution matching. Besides the semi-supervised image classification task considered in [12], we also conduct experiments on image-to-image translation and attribute-conditional image generation. When modeling image pairs, both $p_x(\mathbf{x}|\mathbf{y})$ and $p_y(\mathbf{y}|\mathbf{x})$ are implemented without introducing additional latent variables, *i.e.*, $p_x(\mathbf{x}|\mathbf{y}) = \delta(\mathbf{x} - G_x(\mathbf{y}))$, $p_y(\mathbf{y}|\mathbf{x}) = \delta(\mathbf{y} - G_y(\mathbf{x}))$.

A different strategy is adopted when modeling the image-label/attribute pairs. Specifically, let \mathbf{x} denote samples in the image domain, \mathbf{y} denote samples in the label/attribute domain. \mathbf{y} is a one-hot vector or a binary vector when representing labels and attributes, respectively. When modeling $p_x(\mathbf{x}|\mathbf{y})$, we assume that \mathbf{x} is transformed by the latent style variables \mathbf{z} given the label or attribute vector \mathbf{y} , *i.e.*, $p_x(\mathbf{x}|\mathbf{y}) = \int \delta(\mathbf{x} - G_x(\mathbf{y}, \mathbf{z}))p(\mathbf{z})d\mathbf{z}$, where $p(\mathbf{z})$ is chosen to be a simple distribution (*e.g.*, uniform or standard normal). When learning $p_y(\mathbf{y}|\mathbf{x})$, $p_y(\mathbf{y}|\mathbf{x})$ is assumed to be a standard multi-class or multi-label classifier without latent variables \mathbf{z} . In order to allow the training signal backpropagated from D_1 and D_2 to G_y , instead of sampling a random variable \mathbf{y} , we simply choose the label with the maximum probability as the predicted label, or use the output of the sigmoid function as the predicted attribute vector.

3 Related work

The proposed framework focuses on designing GAN for joint-distribution matching. Conditional GAN can be used for this task if supervised data is available. Various conditional GANs have been proposed to condition the image generation on class labels [6], attributes [15], texts [4, 16] and images [5, 17]. Unsupervised learning methods have also been developed for this task. BiGAN [8] and ALI [7] proposed a method to jointly learn a generation network and an inference network via adversarial learning. Though originally designed for learning the two-way transition between the stochastic latent variables and real data samples, BiGAN and ALI can be directly adapted to learn the joint distribution of two real domains. Another method is called DiscoGAN [9], in which two generators are used to model the bidirectional mapping between domains, and another two discriminators are used to decide whether a generated sample is fake or not in each individual domain. Further, additional reconstruction losses are introduced to make the two generators strongly coupled and also alleviate the problem of mode collapsing. Similar work includes CycleGAN [10], DualGAN [11] and DTN [18]. Additional weight-sharing constraints are introduced in CoGAN [19] and UNIT [20].

Our work differs from the above work in that we aim at semi-supervised joint distribution matching. The only work that we are aware of that also achieves this goal is Triple GAN. However, our model is distinct from Triple GAN in important ways (see Section 2.5). Further, Triple GAN only focuses on the task of image classification, while Δ -GAN has been shown to be applicable to a wide range of applications.

Various methods and model architectures have been proposed to improve and stabilize the training of GAN, such as feature matching [21], Wasserstein GAN [22], energy-based GAN [23], Stein GAN [24] and unrolled GAN [25] among many other related works. Our work is orthogonal to these methods, which could also be used to improve the training of Δ -GAN.

4 Experiments

We present results on three tasks: (*i*) semi-supervised classification on CIFAR10 [26]; (*ii*) image-to-image translation on MNIST [27] and the edges2shoes dataset [5]; and (*iii*) attribute-to-image

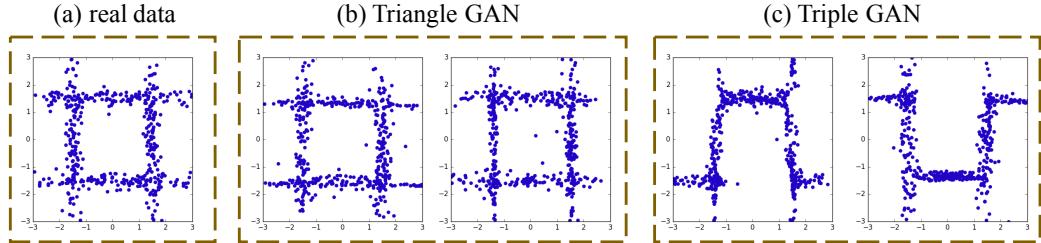


Figure 2: Toy data experiment on Δ -GAN and Triple GAN. (a) the joint distribution $p(x, y)$ of real data. For (b) and (c), the left and right figure is the learned joint distribution $p_x(x, y)$ and $p_y(x, y)$, respectively.

Table 1: Error rates (%) on the partially labeled CIFAR10 dataset.

Algorithm	$n = 4000$
Ladder network [29]	20.40 ± 0.47
CatGAN [30]	19.58 ± 0.58
Improved GAN [21]	18.63 ± 2.32
Triple GAN [12]	18.82 ± 0.32
ALI [7]	17.99 ± 1.62
Δ -GAN (ours)	17.17 ± 0.45

Table 2: Classification accuracy (%) on the MNIST-to-MNIST-transpose dataset.

Algorithm	$n = 100$	$n = 1000$	All
DiscoGAN	—	—	15.00 ± 0.20
Triple GAN	63.79 ± 0.85	84.93 ± 1.63	86.70 ± 1.52
Δ -GAN	83.20 ± 1.88	88.98 ± 1.50	93.34 ± 1.46

generation on CelebA [28] and COCO [14]. We also conduct a toy data experiment to further demonstrate the differences between Δ -GAN and Triple GAN. We implement Δ -GAN without introducing additional regularization unless explicitly stated. All the network architectures are provided in the Appendix.

4.1 Toy data experiment

We first compare our method with Triple GAN on a toy dataset. We synthesize data by drawing $(x, y) \sim \frac{1}{4}\mathcal{N}(\mu_1, \Sigma_1) + \frac{1}{4}\mathcal{N}(\mu_2, \Sigma_2) + \frac{1}{4}\mathcal{N}(\mu_3, \Sigma_3) + \frac{1}{4}\mathcal{N}(\mu_4, \Sigma_4)$, where $\mu_1 = [0, 1.5]^\top$, $\mu_2 = [-1.5, 0]^\top$, $\mu_3 = [1.5, 0]^\top$, $\mu_4 = [0, -1.5]^\top$, $\Sigma_1 = \Sigma_4 = \begin{pmatrix} 3 & 0 \\ 0 & 0.025 \end{pmatrix}$ and $\Sigma_2 = \Sigma_3 = \begin{pmatrix} 0.025 & 0 \\ 0 & 3 \end{pmatrix}$. We generate 5000 (x, y) pairs for each mixture component. We model $p_x(x|y)$ and $p_y(y|x)$ as $p_x(x|y) = \int \delta(x - G_x(y, z))p(z)dz$, $p_y(y|x) = \int \delta(y - G_y(x, z))p(z)dz$ where both G_x and G_y are modeled as a 4-hidden-layer multilayer perceptron (MLP) with 500 hidden units in each layer. $p(z)$ is a bivariate standard Gaussian distribution. Due to the fact that the explicit density functions of $p_x(x|y)$ and $p_y(y|x)$ are not available, Triple GAN cannot be implemented. Hence, we implement Triple GAN-s instead (see Section 2.5).

Results are shown in Figure 2. The joint distributions $p_x(x, y)$ and $p_y(x, y)$ learned by Δ -GAN successfully match the true joint distribution $p(x, y)$. Triple GAN-s cannot achieve this, and can only guarantee $\frac{1}{2}(p_x(x, y) + p_y(x, y))$ matches $p(x, y)$. Although this experiment is limited due to its simplicity, the results clearly support the advantage of our proposed model over Triple GAN.

4.2 Semi-supervised classification

We evaluate semi-supervised classification on the CIFAR10 dataset with 4000 labels. The labeled data is distributed equally across classes and the results are averaged over 10 runs with different random splits of the training data. For fair comparison, we follow the publically available code of Triple GAN and use the same regularization terms and hyperparameter settings as theirs. Results are summarized in Table 1. Our Δ -GAN achieves the best performance among all the competing methods. We also show the ability of Δ -GAN to disentangle classes and styles in Figure 3. Δ -GAN can generate realistic data in a specific class and the injected noise vector encodes meaningful style patterns like background and color.

4.3 Image-to-image translation

We first evaluate image-to-image translation on the edges2shoes dataset. Results are shown in Figure 4(bottom). Though DiscoGAN is an unsupervised learning method, it achieves impressive



Figure 3: Generated CIFAR10 samples, where each row shares the same label and each column uses the same noise.

	DiscoGAN				Δ -GAN			
Input:	4	5	0	7	7	5	9	5
Output:	3	4	8	2	8	4	5	6
Input:	6	9	0	-	8	-	0	4
Output:	4	9	9	0	8	1	0	9
	Shoe	Horse	Boots	Red Boot	Shoe	Horse	Boots	Shoe
GT Output:								
DiscoGAN:								
Δ -GAN:								

Figure 4: Image-to-image translation experiments on the MNIST-to-MNIST-transpose and edges2shoes datasets.



Figure 5: Results on the face-to-attribute-to-face experiment. The 1st row is the input images; the 2nd row is the predicted attributes given the input images; the 3rd row is the generated images given the predicted attributes.

Table 3: Results of P@10 and nDCG@10 for attribute predicting on CelebA and COCO.

Dataset	CelebA			COCO			
	Method	1%	10%	100%	10%	50%	100%
Triple GAN	40.97/50.74	62.13/73.56	70.12/79.37	32.64/35.91	34.00/37.76	35.35/39.60	
Δ -GAN	53.21/58.39	63.68/75.22	70.37/81.47	34.38/37.91	36.72/40.39	39.05/42.86	

results. However, with supervision provided by 10% paired data, Δ -GAN generally generates more accurate edge details of the shoes. In order to provide quantitative evaluation of translating shoes to edges, we use mean squared error (MSE) as our metric. The MSE of using DiscoGAN is 140.1; with 10%, 20%, 100% paired data, the MSE of using Δ -GAN is 125.3, 113.0 and 66.4, respectively.

To further demonstrate the importance of providing supervision of domain correspondence, we created a new dataset based on MNIST [27], where the two image domains are the MNIST images and their corresponding transposed ones. As can be seen in Figure 4(top), Δ -GAN matches images between domains well, while DiscoGAN fails in this task. For supporting quantitative evaluation, we have trained a classifier on the MNIST dataset, and the classification accuracy of this classifier on the test set approaches 99.4%, and is, therefore, trustworthy as an evaluation metric. Given an input MNIST image, we first generate a transposed image using the learned generator, and then transpose it back using the other learned generator, and finally send this new image to the classifier. Results are summarized in Table 2, which are averages over 5 runs with different random splits of the training data. Δ -GAN achieves significantly better performance than Triple GAN and DiscoGAN.

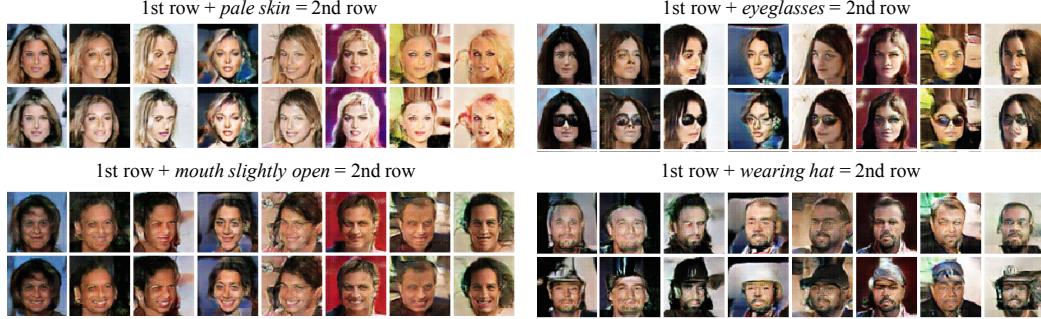


Figure 6: Results on the image editing experiment.

Input	Predicted attributes	Generated images	Input	Predicted attributes	Generated images
	baseball, standing, next, player, man, group, person, field, sport, ball, outdoor, game, grass, crowd			tennis, player, court, man, playing, field, racket, sport, swinging, ball, outdoor, holding, game, grass	
	surfing, people, woman, water, standing, wave, man, top, riding, sport, ocean, outdoor, board			skiing, man, group, covered, day, hill, person, snow, riding, outdoor	
	red, sign, street, next, pole, outdoor, stop, grass			pizza, rack, blue, grill, plate, stove, table, pan, holding, pepperoni, cooked	
	sink, shower, indoor, tub, restroom, bathroom, small, standing, room, tile, white, stall, tiled, black, bath			computer, laptop, room, front, living, indoor, table, desk	

Figure 7: Results on the image-to-attribute-to-image experiment.

4.4 Attribute-conditional image generation

We apply our method to face images from the CelebA dataset. This dataset consists of 202,599 images annotated with 40 binary attributes. We scale and crop the images to 64×64 pixels. In order to qualitatively evaluate the learned attribute-conditional image generator and the multi-label classifier, given an input face image, we first use the classifier to predict attributes, and then use the image generator to produce images based on the predicted attributes. Figure 5 shows example results. Both the learned attribute predictor and the image generator provides good results. We further show another set of image editing experiment in Figure 6. It is interesting to see that Δ -GAN has great flexibility to adjust the generated images by changing certain input attributes. For instance, by switching on the *wearing hat* attribute, one can edit the face image to have a hat on the head.

In order to demonstrate the scalability of our model to large and complex datasets, we also present results on the COCO dataset. Following [31], we first select a set of attributes from the caption text in the training set. We use the 1000 most common words in the training captions to determine the vocabulary of attributes, which includes the most frequent nouns, verbs, or adjectives. The images in COCO are scaled and cropped to have 64×64 pixels. Unlike the case of CelebA face images, the networks need to learn how to handle multiple objects and diverse backgrounds. Results are provided in Figure 7. We can generate reasonably good images based on the predicted attributes. The input and generated images also clearly share a same set of attributes. We also observe diversity in the samples by simply drawing multiple noise vectors and using the same predicted attributes.

Precision (P) and normalized Discounted Cumulative Gain (nDCG) are two popular evaluation metrics for multi-label classification problems. Table 3 provides the quantitative results of P@10 and nDCG@10 on CelebA and COCO, where $@k$ means at rank k (see the Appendix for definitions). For fair comparison, we use the same network architectures for both Triple GAN and Δ -GAN. Δ -GAN consistently provides better results than Triple GAN. On the COCO dataset, our semi-supervised learning approach with 50% labeled data achieves better performance than the results of Triple GAN using the full dataset, demonstrating the effectiveness of our approach for semi-supervised joint distribution matching. More results for the above experiments are provided in the Appendix.

5 Conclusion

We have presented the Triangle Generative Adversarial Network (Δ -GAN), a new GAN framework that can be used for semi-supervised joint distribution matching. Our approach learns the bidirectional mappings between two domains with a few paired samples. We have demonstrated that Δ -GAN may be employed for a wide range of applications. One possible future direction is to combine Δ -GAN and sequence GAN [32] to model the joint distribution of image-text pairs.

References

- [1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.
- [2] Emily Denton, Soumith Chintala, Arthur Szlam, and Rob Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. In *NIPS*, 2015.
- [3] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016.
- [4] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *ICML*, 2016.
- [5] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.
- [6] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv:1411.1784*, 2014.
- [7] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Alex Lamb, Martin Arjovsky, Olivier Mastropietro, and Aaron Courville. Adversarially learned inference. In *ICLR*, 2017.
- [8] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. In *ICLR*, 2017.
- [9] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jungkwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. *arXiv:1703.05192*, 2017.
- [10] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv:1703.10593*, 2017.
- [11] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. *arXiv:1704.02510*, 2017.
- [12] Chongxuan Li, Kun Xu, Jun Zhu, and Bo Zhang. Triple generative adversarial nets. *arXiv:1703.02291*, 2017.
- [13] Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. In *ICLR*, 2017.
- [14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.
- [15] Guim Perarnau, Joost van de Weijer, Bogdan Raducanu, and Jose M Álvarez. Invertible conditional gans for image editing. *arXiv:1611.06355*, 2016.
- [16] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *CVPR*, 2017.
- [17] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv:1609.04802*, 2016.
- [18] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. In *ICLR*, 2017.
- [19] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *NIPS*, 2016.
- [20] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *arXiv:1703.00848*, 2017.
- [21] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NIPS*, 2016.
- [22] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv:1701.07875*, 2017.
- [23] Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. In *ICLR*, 2017.
- [24] Dilin Wang and Qiang Liu. Learning to draw samples: With application to amortized mle for generative adversarial learning. *arXiv:1611.01722*, 2016.
- [25] Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled generative adversarial networks. In *ICLR*, 2017.
- [26] Alex Krizhevsky. Learning multiple layers of features from tiny images. *CiteSeer*, 2009.
- [27] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

- [28] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015.
- [29] Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. In *NIPS*, 2015.
- [30] Jost Tobias Springenberg. Unsupervised and semi-supervised learning with categorical generative adversarial networks. *arXiv:1511.06390*, 2015.
- [31] Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. Semantic compositional networks for visual captioning. In *CVPR*, 2017.
- [32] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: sequence generative adversarial nets with policy gradient. In *AAAI*, 2017.

A Detailed theoretical analysis

Proposition 3. For any fixed generator G_x and G_y , the optimal discriminator D_1 and D_2 of the game defined by the value function $V(G_x, G_y, D_1, D_2)$ is

$$D_1^*(\mathbf{x}, \mathbf{y}) = \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x}, \mathbf{y}) + p_x(\mathbf{x}, \mathbf{y}) + p_y(\mathbf{x}, \mathbf{y})}, \quad D_2^*(\mathbf{x}, \mathbf{y}) = \frac{p_x(\mathbf{x}, \mathbf{y})}{p_x(\mathbf{x}, \mathbf{y}) + p_y(\mathbf{x}, \mathbf{y})}$$

Proof. The training criterion for the discriminator D_1 and D_2 , given any generator G_x and G_y , is to maximize the quantity $V(G_x, G_y, D_1, D_2)$:

$$\begin{aligned} V(G_x, G_y, D_1, D_2) &= \int_{\mathbf{x}} \int_{\mathbf{y}} p(\mathbf{x}, \mathbf{y}) \log D_1(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} + \int_{\mathbf{x}} \int_{\mathbf{y}} p_x(\mathbf{x}, \mathbf{y}) \log(1 - D_1(\mathbf{x}, \mathbf{y})) d\mathbf{x} d\mathbf{y} \\ &\quad + \int_{\mathbf{x}} \int_{\mathbf{y}} p_x(\mathbf{x}, \mathbf{y}) \log D_2(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} + \int_{\mathbf{x}} \int_{\mathbf{y}} p_y(\mathbf{x}, \mathbf{y}) \log(1 - D_1(\mathbf{x}, \mathbf{y})) d\mathbf{x} d\mathbf{y} \\ &\quad + \int_{\mathbf{x}} \int_{\mathbf{y}} p_y(\mathbf{x}, \mathbf{y}) \log(1 - D_2(\mathbf{x}, \mathbf{y})) d\mathbf{x} d\mathbf{y} \end{aligned}$$

Following [1], for any $(a, b) \in \mathbb{R}^2 \setminus \{0, 0\}$, the function $y \rightarrow a \log y + b \log(1 - y)$ achieves its maximum in $[0, 1]$ at $\frac{a}{a+b}$. This concludes the proof. \square

Proposition 4. The equilibrium of $V(G_x, G_y, D_1, D_2)$ is achieved if and only if $p(\mathbf{x}, \mathbf{y}) = p_x(\mathbf{x}, \mathbf{y}) = p_y(\mathbf{x}, \mathbf{y})$ with $D_1^*(\mathbf{x}, \mathbf{y}) = \frac{1}{3}$ and $D_2^*(\mathbf{x}, \mathbf{y}) = \frac{1}{2}$, and the optimum value is $-3 \log 3$.

Proof. Given the optimal $D_1^*(\mathbf{x}, \mathbf{y})$ and $D_2^*(\mathbf{x}, \mathbf{y})$, the minimax game can be reformulated as:

$$C(G_x, G_y) = \max_{D_1, D_2} V(G_x, G_y, D_1, D_2) \tag{9}$$

$$= \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim p(\mathbf{x}, \mathbf{y})} \left[\log \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x}, \mathbf{y}) + p_x(\mathbf{x}, \mathbf{y}) + p_y(\mathbf{x}, \mathbf{y})} \right] \tag{10}$$

$$+ \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim p_x(\mathbf{x}, \mathbf{y})} \left[\log \frac{p_x(\mathbf{x}, \mathbf{y})}{p(\mathbf{x}, \mathbf{y}) + p_x(\mathbf{x}, \mathbf{y}) + p_y(\mathbf{x}, \mathbf{y})} \right] \tag{11}$$

$$+ \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim p_y(\mathbf{x}, \mathbf{y})} \left[\log \frac{p_y(\mathbf{x}, \mathbf{y})}{p(\mathbf{x}, \mathbf{y}) + p_x(\mathbf{x}, \mathbf{y}) + p_y(\mathbf{x}, \mathbf{y})} \right] \tag{12}$$

Note that

$$C(G_1, G_2) = -3 \log 3 + KL \left(p(\mathbf{x}, \mathbf{y}) \middle\| \frac{p(\mathbf{x}, \mathbf{y}) + p_x(\mathbf{x}, \mathbf{y}) + p_y(\mathbf{x}, \mathbf{y})}{3} \right) \tag{13}$$

$$+ KL \left(p_x(\mathbf{x}, \mathbf{y}) \middle\| \frac{p(\mathbf{x}, \mathbf{y}) + p_x(\mathbf{x}, \mathbf{y}) + p_y(\mathbf{x}, \mathbf{y})}{3} \right) \tag{14}$$

$$+ KL \left(p_y(\mathbf{x}, \mathbf{y}) \middle\| \frac{p(\mathbf{x}, \mathbf{y}) + p_x(\mathbf{x}, \mathbf{y}) + p_y(\mathbf{x}, \mathbf{y})}{3} \right) \tag{15}$$

Therefore,

$$C(G_1, G_2) = -3 \log 3 + 3 \cdot JSD \left(p(\mathbf{x}, \mathbf{y}), p_x(\mathbf{x}, \mathbf{y}), p_y(\mathbf{x}, \mathbf{y}) \right) \geq -3 \log 3 \tag{16}$$

where $JSD_{\pi_1, \dots, \pi_n}(p_1, p_2, \dots, p_n) = H \left(\sum_{i=1}^n \pi_i p_i \right) - \sum_{i=1}^n \pi_i H(p_i)$ is the Jensen-Shannon divergence.

π_1, \dots, π_n are weights that are selected for the probability distribution p_1, p_2, \dots, p_n , and $H(p)$ is the entropy for distribution p . In the three-distribution case described above, we set $n = 3$ and $\pi_1 = \pi_2 = \pi_3 = \frac{1}{3}$.



Figure 8: Additional results on the image-to-image translation experiment.

For $p(\mathbf{x}, \mathbf{y}) = p_x(\mathbf{x}, \mathbf{y}) = p_y(\mathbf{x}, \mathbf{y})$, we have $D_1^*(\mathbf{x}, \mathbf{y}) = \frac{1}{3}$, $D_2^*(\mathbf{x}, \mathbf{y}) = \frac{1}{2}$ and $C(G_x, G_y) = -3 \log 3$. Since the Jensen-Shannon divergence is always non-negative, and zero iff they are equal, we have shown that $C^* = -3 \log 3$ is the global minimum of $C(G_x, G_y)$ and that the only solution is $p(\mathbf{x}, \mathbf{y}) = p_x(\mathbf{x}, \mathbf{y}) = p_y(\mathbf{x}, \mathbf{y})$, i.e., the generative models perfectly replicating the data distribution. \square

B Δ -GAN training procedure

Algorithm 1 Δ -GAN training procedure.

```

 $\theta_g, \theta_d \leftarrow$  initialize network parameters
repeat
     $(\mathbf{x}_p^{(1)}, \mathbf{y}_p^{(1)}), \dots, (\mathbf{x}_p^{(M)}, \mathbf{y}_p^{(M)}) \sim p(\mathbf{x}, \mathbf{y})$                                  $\triangleright$  Get  $M$  paired data samples
     $\mathbf{x}_u^{(1)}, \dots, \mathbf{x}_u^{(M)} \sim p(\mathbf{x})$                                                $\triangleright$  Get  $M$  unpaired data samples
     $\mathbf{y}_u^{(1)}, \dots, \mathbf{y}_u^{(M)} \sim p(\mathbf{y})$ 
     $\tilde{\mathbf{x}}_u^{(i)} \sim p_x(\mathbf{x}|\mathbf{y} = \mathbf{y}_u^{(i)}), \quad i = 1, \dots, M$                                  $\triangleright$  Sample from the conditionals
     $\tilde{\mathbf{y}}_u^{(j)} \sim p_y(\mathbf{y}|\mathbf{x} = \mathbf{x}_u^{(j)}), \quad j = 1, \dots, M$ 
     $\rho_{11}^{(i)} \leftarrow D_1(\mathbf{x}_p^{(i)}, \mathbf{y}_p^{(i)}), \quad i = 1, \dots, M$                                  $\triangleright$  Compute discriminator predictions
     $\rho_{12}^{(i)} \leftarrow D_1(\tilde{\mathbf{x}}_u^{(i)}, \mathbf{y}_p^{(i)}), \quad i = 1, \dots, M$ 
     $\rho_{13}^{(i)} \leftarrow D_1(\mathbf{x}_p^{(i)}, \tilde{\mathbf{y}}_u^{(i)}), \quad i = 1, \dots, M$ 
     $\rho_{21}^{(i)} \leftarrow D_2(\tilde{\mathbf{x}}_p^{(i)}, \mathbf{y}_p^{(i)}), \quad i = 1, \dots, M$ 
     $\rho_{22}^{(i)} \leftarrow D_2(\mathbf{x}_p^{(i)}, \tilde{\mathbf{y}}_p^{(i)}), \quad i = 1, \dots, M$ 
     $\mathcal{L}_{d_1} \leftarrow -\frac{1}{M} \sum_{i=1}^M \log \rho_{11}^{(i)} - \frac{1}{M} \sum_{j=1}^M \log(1 - \rho_{12}^{(j)}) - \frac{1}{M} \sum_{k=1}^M \log(1 - \rho_{13}^{(k)})$ 
     $\mathcal{L}_{d_2} \leftarrow -\frac{1}{M} \sum_{i=1}^M \log \rho_{21}^{(i)} - \frac{1}{M} \sum_{j=1}^M \log(1 - \rho_{22}^{(j)})$                                  $\triangleright$  Compute discriminator loss
     $\mathcal{L}_{g_1} \leftarrow -\frac{1}{M} \sum_{i=1}^M \log \rho_{12}^{(i)} - \frac{1}{M} \sum_{j=1}^M \log(1 - \rho_{21}^{(j)})$                                  $\triangleright$  Compute generator loss
     $\mathcal{L}_{g_2} \leftarrow -\frac{1}{M} \sum_{i=1}^M \log \rho_{13}^{(i)} - \frac{1}{M} \sum_{j=1}^M \log \rho_{22}^{(j)}$ 
     $\theta_d \leftarrow \theta_d - \nabla_{\theta_d} (\mathcal{L}_{d_1} + \mathcal{L}_{d_2})$                                  $\triangleright$  Gradient update on discriminator networks
     $\theta_g \leftarrow \theta_g - \nabla_{\theta_g} (\mathcal{L}_{g_1} + \mathcal{L}_{g_2})$                                  $\triangleright$  Gradient update on generator networks
until convergence

```

C Additional experimental results



Figure 9: Generated CelebA samples, where each row shares the same set of attributes and each column uses the same noise.

Input images								
Predicted attributes	Attractive, Smiling, High Cheekbones, No Beard, Oval Face, Young	Smiling, Mouth Slightly Open, Rosy Cheeks, Wearing Lipstick	Black Hair, No Beard, Young, Wearing Lipstick	Bags Under Eyes, No Beard, Pointy Nose, Smiling, Wearing Lipstick	Blond Hair, Mouth Slightly Open, Narrow Eyes, No Beard, Straight Hair, Young	Attractive, Black Hair, Male, No Beard, Pointy Nose, Straight Hair	Attractive, Blond Hair, No Beard, Pointy Nose, Smiling, Straight Hair	Brown Hair, Bushy Eyebrows, High Cheekbones, Young
Generated images								

Figure 10: Additional results on the face-to-attribute-to-face experiment.

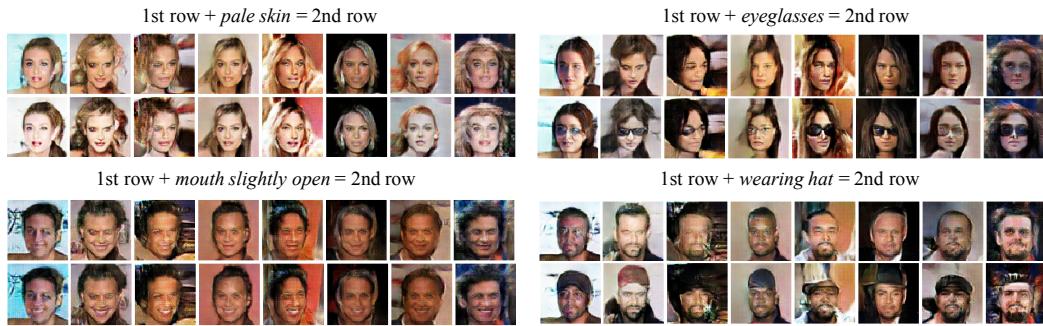


Figure 11: Additional results on the image editing experiment.

Input	Predicted attributes	Generated images	Input	Predicted attributes	Generated images
	Building, standing, tall, castle, city, top, object, outdoor, tower			airport, airplane, cloudy, large, tarmac, parked, jet, commercial, white, gear, plane, field, flying, landing, aircraft, runway, air, transport	
	furniture, sitting, small, room, living, white, hotel, indoor, table, photo, rug, decorated, window, cabinet			mammal, standing, animal, field, walking, outdoor, grass	
	kite, people, young, blue, boy, standing, playing, colorful, child, air, outdoor, holding, girl, flying			large, red, street, parking, standing, next, decker, tall, train, parked, city, outdoor, transport, tour, road	

Figure 12: Additional results on the image-to-attribute-to-image experiment.

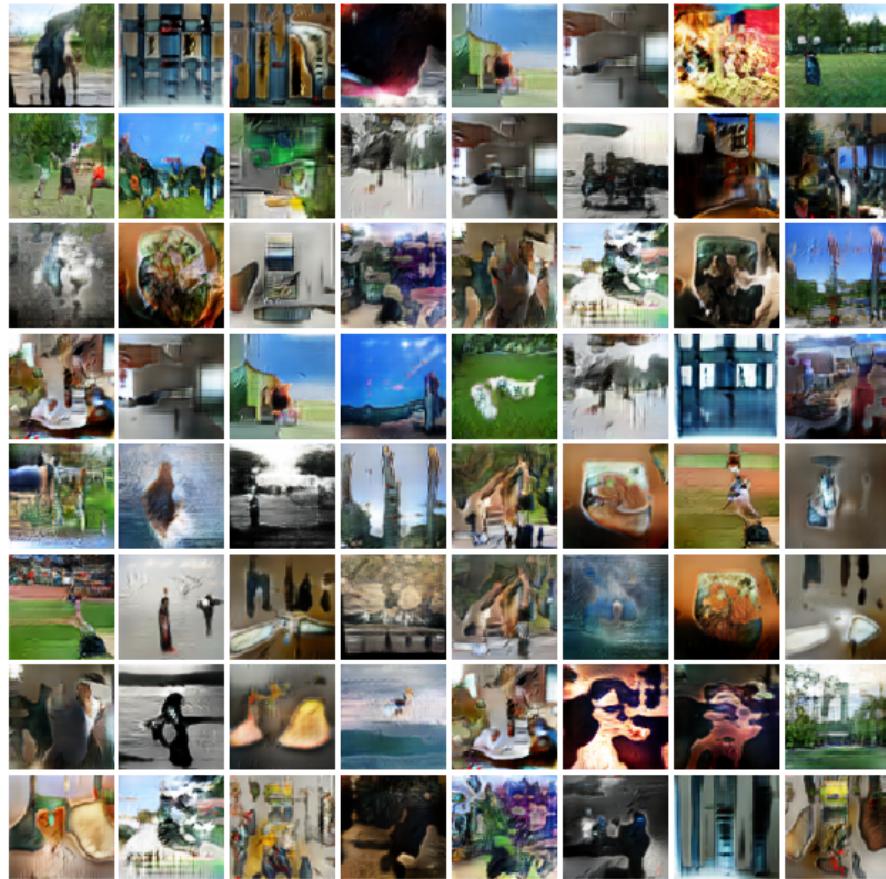


Figure 13: Attribute-conditional image generation on the COCO dataset. Input attributes are omitted for brevity.

Table 4: Results of P@5 and nDCG@5 for attribute predicting on CelebA and COCO.

Dataset		CelebA			COCO		
Method	1%	10%	100%	10%	50%	100%	
Triple GAN	49.35/52.73	73.68/74.55	80.89/78.58	38.98/41.00	41.08/43.50	43.18/46.00	
Δ -GAN	59.55/60.53	74.06/75.49	80.39/79.41	41.51/43.55	44.42/46.40	47.32/49.24	

Table 5: Results of P@3 and nDCG@3 for attribute predicting on CelebA and COCO.

Dataset		CelebA			COCO		
Method	1%	10%	100%	10%	50%	100%	
Triple GAN	55.30/56.87	76.09/75.44	83.54/84.74	42.23/43.60	45.35/46.85	48.47/50.10	
Δ -GAN	62.62/62.72	76.04/76.27	84.81/86.85	45.45/46.56	48.19/49.29	50.92/52.02	

D Evaluation metrics for multi-label classification

Precision@ k Precision at k is a popular evaluation metric for multi-label classification problems. Given the ground truth label vector $\mathbf{y} \in \{0, 1\}^L$ and the prediction $\hat{\mathbf{y}} \in [0, 1]^L$, $P@k$ is defined as

$$P@k := \frac{1}{k} \sum_{l \in \text{rank}_k(\hat{\mathbf{y}})} y^{(l)}.$$

Precision at k performs a sentence-wise evaluation that counts the fraction of correct predictions in the top k scoring labels.

nDCG@ k normalized Discounted Cumulative Gain (nDCG) at rank k is a family of ranking measures widely used in multi-label learning. DCG is the total gain accumulated at a particular rank p , which is defined as

$$DCG@k := \sum_{l \in \text{rank}_k(\hat{\mathbf{y}})} \frac{y^{(l)}}{\log(l+1)}.$$

Then normalizing DCG by the value at rank k of the ideal ranking gives

$$N@k := \frac{DCG@k}{\sum_{l=1}^{\min(k, \|\mathbf{y}\|_0)} \frac{1}{\log(l+1)}}$$

Here, nDCG at k ($N@k$) performs a sentence-wise evaluation.

E Detailed network architectures

For the CIFAR10 dataset, we use the same network architecture as used in Triple GAN [12]. For the edges2shoes dataset, we use the same network architecture as used in the pix2pix paper [5]. For other datasets, we provide the detailed network architectures below.

Table 6: Architecture of the models for Δ -GAN on MNIST. BN denotes batch normalization.

Generator A to B		Generator B to A		Discriminator
Input 28 × 28 Gray Image	Input 28 × 28 Gray Image	Input 28 × 28 Gray Image	Input two 28 × 28 Gray Image	
5 × 5 conv. 32 ReLU, stride 2, BN	5 × 5 conv. 32 ReLU, stride 2, BN	5 × 5 conv. 32 ReLU, stride 2, BN	5 × 5 conv. 32 ReLU, stride 2, BN	
5 × 5 conv. 64 ReLU, stride 2, BN	5 × 5 conv. 64 ReLU, stride 2, BN	5 × 5 conv. 64 ReLU, stride 2, BN	5 × 5 conv. 64 ReLU, stride 2, BN	
5 × 5 conv. 128 ReLU, stride 2, BN	5 × 5 conv. 128 ReLU, stride 2, BN	5 × 5 conv. 128 ReLU, stride 2, BN	5 × 5 conv. 128 ReLU, stride 2, BN	
Dropout: 0.1	Dropout: 0.1	Dropout: 0.1	Dropout: 0.1	
MLP output 28 × 28, sigmoid	MLP output 28 × 28, sigmoid	MLP output 28 × 28, sigmoid	MLP output 1, sigmoid	

Table 7: Architecture of the models for Δ -GAN on CelebA. BN denotes batch normalization. IReLU denotes Leaky ReLU.

Generator A to B	Generator B to A	Discriminator
Input $64 \times 64 \times 3$ Image	Input 1×40 attributes, 1×100 noise	Input 64×64 Image and 1×40 attributes
4×4 conv. 32 IReLU, stride 2, BN	concat input	concat two inputs
4×4 conv. 64 IReLU, stride 2, BN	MLP output 1024, IReLU, BN	5×5 conv. 64 ReLU, stride 2, BN
4×4 conv. 128 IReLU, stride 2, BN	MP output 8192, IReLU, BN	5×5 conv. 128 ReLU, stride 2, BN
4×4 conv. 256 IReLU, stride 2, BN	concat attributes	
4×4 conv. 512 IReLU, stride 2, BN	5×5 deconv. 256 ReLU, stride 2, BN	5×5 conv. 256 ReLU, stride 2, B
MLP output 512, IReLU	5×5 deconv. 128 ReLU, stride 2, BN	5×5 conv. 512 ReLU, stride 2, BN
MLP output 40, sigmoid	5×5 deconv. 64 ReLU, stride 2, BN	
	5×5 deconv. 3 tanh, stride 2, BN	MLP output 1, sigmoid

Table 8: Architecture of the models for Δ -GAN on COCO. BN denotes batch normalization. IReLU denotes Leaky ReLU. Dim denotes the number of attributes.

Generator A to B	Generator B to A	Discriminator
Input $64 \times 64 \times 3$ Image	Input 1×40 attributes, 1×100 noise	Input 64×64 Image and $1 \times Dim$ attributes
4×4 conv. 32 IReLU, stride 2, BN	concat inputs	concat conditional inputs
4×4 conv. 64 IReLU, stride 2, BN	MLP output 16384, BN	
4×4 conv. 128 IReLU, stride 2, BN	ResNet Block	5×5 conv. 64 ReLU, stride 2, BN
4×4 conv. 256 IReLU, stride 2, BN	4×4 deconv. 512, stride 2	
4×4 conv. 512 IReLU, stride 2, BN	3×3 conv. 512, stride 1, BN	5×5 conv. 128 ReLU, stride 2, BN
ResNet Block	ResNet Block	
	4×4 deconv. 256, stride 2	5×5 conv. 256 ReLU, stride 2, BN
	3×3 conv. 256, stride 1, BN	
	4×4 deconv. 128 ReLU, stride 2	5×5 conv. 512 ReLU, stride 2, BN
1×1 conv. 512 IReLU, stride 1, BN	3×3 conv. 128 ReLU, stride 1, BN	
4×4 conv. Dim sigmoid, stride 4	4×4 deconv. Dim , stride 2	MLP output 1, sigmoid
	3×3 conv. Dim tanh, stride 1	