Learning Energy-Based Models With Adversarial Training

Xuwang Yin, Shiying Li, and Gustavo K. Rohde

University of Virginia
{xy4cm, sl8jx, gustavo}@virginia.edu

Abstract. We study a new approach to learning energy-based models (EBMs) based on adversarial training (AT). We show that (binary) AT learns a special kind of energy function that models the support of the data distribution, and the learning process is closely related to MCMC-based maximum likelihood learning of EBMs. We further propose improved techniques for generative modeling with AT, and demonstrate that this new approach is capable of generating diverse and realistic images. Aside from having competitive image generation performance to explicit EBMs, the studied approach is stable to train, is well-suited for image translation tasks, and exhibits strong out-of-distribution adversarial robustness. Our results demonstrate the viability of the AT approach to generative modeling, suggesting that AT is a competitive alternative approach to learning EBMs.

Keywords: Adversarial Training, Adversarial Attacks, Energy-Based Models (EBMs), Generative Modeling

1 Introduction

In unsupervised learning, energy-based models (EBMs) [33] are a class of generative model that uses an energy function to model the probability distribution of the observed data. Unlike explicit density models, EBMs model the unnormalized density function, which makes it difficult to evaluate the likelihood function. Maximum likelihood learning of EBMs hence makes use of the likelihood function's gradient which can be approximated using Monte Carlo methods. Each iteration of the learning process involves first generating synthesized data by sampling from the current model, and then updating the model to maximize the energy difference between synthesized data and observed data. This process leads to an energy function that outputs low energies on the data manifold and high energies on other regions. EBMs find applications in image restoration (denoising, inpainting, etc), out-of-distribution detection, and various sample generation tasks. The main difficulties of training EBMs lie in the computational challenges from the sampling procedure and some training stability issues [10,40,39,14,15,9,67,57].

Another line of work on adversarial training (AT) show that adversarially robust classifiers learn high-level, interpretable features, and can be utilized to solve

various computer vision tasks including generation, inpainting, super-resolution, and image-to-image translation [53,25,11,47]. Compared to state-of-the-art generative models, this AT approach does not provide a competitive generation performance and is therefore of limited value in many of these tasks. Nonetheless, the generative properties of the robust classifier suggest that the model has captured the distribution of the training data, although the underlying learning mechanism is not yet well understood.

At a high level, both EBMs training and AT are based on the idea of first using gradient-based optimization to generate samples that reach high activation under the current model, and then optimizing the model to minimize its activation on the generated samples. In addition, both approaches synthesize new samples by performing gradient descent on the trained model. These similarities suggest that there are some connections between these two approaches.

In this work we investigate the mechanism by which AT learns data distributions, and propose improved techniques for generative modeling with AT. We focus on binary AT [65] which does not requires class labels and hence naturally fits the generative modeling task. We first analyze the binary AT objective and the corresponding training algorithm, and show that binary AT learns a special kind of energy function that models the support of the observed data. We then draw a connection between AT and MCMC-base maximum likelihood learning of EBMs by showing that the binary AT objective can be interpreted as a gradient-scaled version of the likelihood objective in EBMs training, and the PGD attack can be viewed as an non-convergent sampler of the model distribution. This connection provides us with intuition of how AT learns data distributions from a maximum likelihood learning perspective, and suggests that binary AT can be viewed as an approximate maximum likelihood learning algorithm.

We further propose improved techniques for generative modeling with AT based on the above analysis. Our empirical evaluation shows that this AT approach provides competitive generation performance to explicit EBMs, and at the same time is stable to train (just like regular adversarial training), is well-suited for image translation tasks, and exhibits strong out-of-distribution adversarial robustness. The main limitation of the studied approach is that it cannot properly learn the underlying density function of the observed data. However, this problem is not unique to the studied approach - most existing work on learning EBMs relies on short-run non-convergent sampler to improve the training efficiency, and the learned model typically does not have a valid steady-state that reflects the distribution of the observed data [40,39].

In summary, the contributions of this paper are: 1) We show that binary AT learns a special kind of energy function that models the support of the data distribution, and the learning process is closely related to MCMC-based maximum likelihood learning of EBMs. 2) We propose improved techniques for generative modeling with AT, and demonstrate competitive image generation performance to state-of-the-art explicit EBMs. 3) We show that the studied approach is stable to train, has competitive training and test time sampling

efficiency, and can be applied to denoising, inpainting, image translation, and worst-case out-of-distribution detection.

2 Related Work

Learning EBMs. Due to the intractability of the normalizing constant in the EBMs likelihood function, maximum likelihood learning of EBMs makes use of the gradient of the log-likelihood which can be approximated using MCMC sampling. Recent work [59,10,40,39] scaling EBMs training to high-dimensional data performs sampling using SGLD [56] and initialize the chain from a noise distribution. The sampling process involves estimating the model's gradient with respect to the current sample at each step and therefore has high computational cost. To improve the sampling efficiency, many authors consider short-run non-convergent SGLD sampler in combination with a persistent sampling buffer [10,39,14,15,57,9]. Although a short-run sampler is sufficient for learning a generation model, the resulting energy function typically does not have a valid steady-state [40,39]. The mixing time of the sampling procedure also depends on how close the chain-initialization distribution is to the model distribution. A recent trend hence considers initializing the sampling chain from samples produced by a generator fitted on the target distribution [58,31,18,41,19,15,57,63,1,38,64].

Maximum likelihood learning of EBMs also has some training stability issues, and various techniques have been developed to address these issues. These techniques include 1) using weight normalization [46], Swish activation [43], gradient clipping, and weight decay (see [57]), 2) gradient norm clipping on model parameters and using a KL term in the training objective (see [9]), 3) adjusting learning rate and SGLD steps during training and adding Gaussian noise to input images (see [14]), 4) gradient clipping on SGLD and model parameters and spectral normalization (see [10]), and 5) multiscale training and smooth activation functions (see [67]). Overall, there does not seem to have a consensus on how to stabilize EBMs training. Due to the computational challenge of MCMC sampling and stability issues, the successful application of EBMs to modeling high-dimensional data such as 256×256 images is only achieved in some very recent works [57,67].

Aside from MCMC-based maximum likelihood learning of EBMs, alternative approaches for learning EBMs exist. Score matching [24] circumvents the difficulty of estimating the partition function by directly modeling the derivatives of the data distribution. Score matching has recently been successfully applied to modeling large natural images and achieves competitive performance to state-of-the-art generative models such as GANs [49,50,51,23]. Noise contrastive estimation (NCE) [17] learns data distributions by contrasting the observed data with data from a known noise distribution. Similar to our approach, NCE makes use of a logistic regression model. The main difference is that in NCE, the logit of the classifier is the difference in log probabilities of the model distribution and the noise distribution, whereas in our approach the logit directly defines the

estimator (i.e., the energy function). Unlike other EBMs, NCE typically does not scale well to high-dimensional data [17,5,44].

Maximin interpretation of EBMs. When the noise term in the SGLD sampler is disabled, the learning process of EBMs can be interpreted as solving a *maximin* game [61,62,60]. This interpretation coincides with our formulation in Eq. (12). The key differences lie in the value function, the setting of the sampler (SGLD vs. PGD attack), and the Markov chain initiation distribution.

3 Background

3.1 Energy-Based Models

Energy-based models (EBMs) [33] represent probability distributions by converting the outputs of a scalar function f_θ into probabilities through a Gibbs distribution:

$$p_\theta(x) = \frac{\exp(f_\theta(x))}{Z(\theta)}, \tag{1}$$

where the normalizing constant $Z(\theta)$, also known as the partition function, is an integral over the unnormalized probability of all states: $Z(\theta) = \int \exp(f_\theta(x))dx$. The energy function is defined as $E_\theta(x) = -f_\theta(x)$, and thus has the property of attributing low energy outputs on the support of the target data distribution and high energy outputs in other regions.

For many interesting models, the partition function $Z(\theta)$ is intractable, and therefore maximum likelihood estimation (MLE) of the model parameters θ is not directly applicable. Standard maximum likelihood learning of EBMs makes use of the gradient of the log likelihood function. Denote the distribution of the observed data as p_{data}, the gradient of the log likelihood takes the form

$$\nabla_\theta \mathbb{E}_{\mathrm{x}\sim p_{\text{data}}}[\log p_\theta(x)] = \mathbb{E}_{\mathrm{x}\sim p_{\text{data}}}[\nabla_\theta f_\theta(x)] - \mathbb{E}_{\mathrm{x}\sim p_\theta(x)}[\nabla_\theta f_\theta(x)]. \tag{2}$$

Intuitively, maximizing log-likelihood with this gradient causes $f_\theta(x)$ to increase on p_{data} samples and decrease on samples drawn from p_θ; when p_θ matches p_{data}, the gradient cancels out and the training terminates.

Evaluating $\mathbb{E}_{\mathrm{x}\sim p_\theta(x)}\nabla_\theta f_\theta(x)$ requires sampling from the model distribution. This can be done with Markov chain Monte Carlo (MCMC) methods. Recent work scaling EBMs training to high-dimensional data [59,10,40,39] makes use of the SGLD method [56] which samples the model distribution by

$$x_0 \sim p_0, \quad x_{i+1} = x_i + \frac{\lambda}{2}\nabla_x f_\theta(x_i) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \lambda), \tag{3}$$

where p_0 is some random noise distribution. A proper SGLD sampler requires a large number of update steps in order for the distribution of sampled data to match p_θ. Due to the high computational cost of this sampling process, many authors resort to short-run non-convergent MCMC to improve the sampling efficiency[40,39,10,59,14]. The resulting model typically does not have a valid steady-state that reflects the distribution of the observed data, but is still capable of generating realistic and diverse samples [40,39].

3.2 Binary Adversarial Training

Binary adversarial training [65] is a method for detecting adversarial examples. In a K class classification problem, the detection method consists of K binary classifiers, with the k-th binary classifier trained to distinguish clean data of class k from adversarially perturbed data of other classes. A committee of K binary classifiers then provides a complete solution for detecting adversarially perturbed samples of any classes.

Denote the data distribution of class k as p_{data}, the mixture distribution of other classes as $p_0 = \frac{1}{K-1}\sum_{i=1,\dots,K,i\neq k} p_i$, the k-th binary classifier is trained by maximizing the objective

$$J(D) = \mathbb{E}_{\mathrm{x}\sim p_{\text{data}}}[\log D(x)] + \mathbb{E}_{\mathrm{x}\sim p_0}[\min_{x'\in\mathbb{B}(x,\epsilon)} \log(1 - D(x'))], \tag{4}$$

where $D : \mathcal{X} \subseteq \mathbb{R}^d \to [0, 1]$ is the classification function, and $\mathbb{B}(x, \epsilon)$ is a neighborhood of x: $\mathbb{B}(x, \epsilon) = \{x' \in \mathcal{X} : \|x' - x\|_2 \leq \epsilon\}$. In practice, D is defined by applying a logistic sigmoid function to the output of a neural network:

$$D(x) = \sigma(f_\theta(x)), \tag{5}$$

where f_θ is a neural network with a single output node and parameters θ.

The inner minimization in Eq. (4) is solved using the PGD attack [34,32], a first-order method that employs an iterative update rule of (l^2-based attack):

$$x_0 \sim p_0, \quad x_{i+1} = \text{Proj}(x_i - \lambda \frac{\nabla_x \log(1 - D(x_i))}{\|\nabla_x \log(1 - D(x_i))\|_2}), \tag{6}$$

where λ is some step size, and Proj is the operation of projecting onto the feasible set $\mathbb{B}(x, \epsilon)$. Because the gradient vector in Eq. (6) is normalized to have unit norm, we can equivalently implement the attack by directly performing gradient ascent on f_θ:

$$x_0 \sim p_0, \quad x_{i+1} = \text{Proj}(x_i + \lambda \frac{\nabla_x f_\theta(x_i)}{\|\nabla_x f_\theta(x_i)\|_2}). \tag{7}$$

4 Binary AT Generative Model

In this section we develop a generative model based on binary AT. We first analyze the optimal solution to the binary AT problem, and then investigate the mechanism by which binary AT learns the data distribution, and finally interpret the learning process from the maximum likelihood learning perspective. Our main result is that under a proper configuration of perturbation limit and p_0 data, binary AT learns a special kind of energy function that models the support of p_{data}. Based on these theoretical insights, we proposed improved training techniques.

4.1 Optimal Solution to the Binary AT Problem

We consider the optimal solution of Eq. (4) under the scenario of unbounded perturbation: $\mathbb{B}(x, \epsilon) = \mathcal{X}$. This allows us to further simplify the PGD attack by removing the Proj operator:

$$x_0 \sim p_0, \quad x_{i+1} = x_i + \lambda \frac{\nabla_x f_\theta(x_i)}{\|\nabla_x f_\theta(x_i)\|_2}. \tag{8}$$

Perturbing p_0 samples can be thought of as moving p_0 samples via a translation function $T(x) = x + \Delta_x$, with Δ_x being the perturbation computed on sample x. We can write the density function of the perturbed distribution p_T using random variable transformation:

$$p_T(y) = \int_{\mathcal{X}} p_0(x)\delta(y - T(x))dx. \tag{9}$$

The inner problem in Eq. (4) can then be interpreted as determining the distribution which has the lowest expected value of $\log(1 - D(x))$:

$$p_T^* = \arg\min_{p_T} \mathbb{E}_{\mathrm{x}\sim p_T}[\log(1 - D(x))]. \tag{10}$$

The objective of the outer problem is then the log-likelihood in a logistic regression model which discriminates p_{data} samples from p_T^* samples:

$$J(D) = \mathbb{E}_{\mathrm{x}\sim p_{\text{data}}}[\log D(x)] + \mathbb{E}_{\mathrm{x}\sim p_T^*}[\log(1 - D(x))]. \tag{11}$$

We can equivalently formulate Eq. (4) as a *maximin* problem

$$\max_D \min_{p_T} U(D, p_T) = \mathbb{E}_{\mathrm{x}\sim p_{\text{data}}}[\log D(x)] + \mathbb{E}_{\mathrm{x}\sim p_T}[\log(1 - D(x))], \tag{12}$$

and obtain its optimal solution by following the standard approach to solving maximin problems:

Proposition 1. *The optimal solution of* $\max_D \min_{p_T} U(D, p_T)$ *is* $U(D^*, p_T^*) = -\log(4)$*, where* D^* *outputs* $\frac{1}{2}$ *on Supp*(p_{data}) *and* $\leq \frac{1}{2}$ *outside Supp*(p_{data})*, and* p_T^* *is supported in the contour set* $\{D = \frac{1}{2}\}$*.*

Proof. See the supplementary materials.

The above maximin problem can also be interpreted as a two-player zero-sum game, and is closely related to GANs [13]'s *minimax* game which has the form

$$\min_G \max_D V(D, G) = \mathbb{E}_{\mathrm{x}\sim p_{\text{data}}}[\log D(x)] + \mathbb{E}_{\mathrm{z}\sim p_z}[\log(1 - D(G(z)))]. \tag{13}$$

The game-theory point of view provides a convenient way to understand their differences. We include a game theory-based analysis of $\max_D \min_{p_T} U(D, p_T)$ and a comparative analysis of GANs in the supplementary materials.

4.2 Learning Mechanism

Proposition 1 states that by solving $\max_D \min_{p_T} U(D, p_T)$ we can obtain a D that outputs $\frac{1}{2}$ on the support of p_{data} and $\leq \frac{1}{2}$ on other regions. This result is obtained by assuming that for any D, the inner minimization Eq. (10) is always perfectly solved. In practice, when D is randomly initialized, it has many local maxima outside the support of p_{data}. Because the inner minimization is solved by taking p_0 samples and then performing gradient ascent on D with Eq. (8), this process can get trapped in different local maxima of D. Hence we can think of this process as searching for these local maxima and then put the perturbed p_0 data in these regions. Then in the model update stage (outer maximization), D is updated by increasing its outputs on p_{data} samples and decreasing its outputs on the perturbed p_0 data. By repeating this process, local maxima get suppressed and the model learns to correctly model $\text{Supp}(p_{\text{data}})$.

The algorithm for solving the maximin problem is described in Algorithm 1. Fig. 1 left panel shows the 2D simulation result of the algorithm when the p_0 dataset contains random samples from the uniform distribution. It can be seen that when the algorithm converges, local maxima outside $\text{Supp}(p_{\text{data}})$ are suppressed, and D (approximately) outputs $\frac{1}{2}$ on $\text{Supp}(p_{\text{data}})$ as predicted by Proposition 1. Meanwhile, D retains the gradient information for translating out-distribution samples to $\text{Supp}(p_{\text{data}})$.

Because the PGD attack is deterministic gradient ascent, its ability to discover different local maxima depends on the diversity of p_0 samples. Fig. 1 right panel shows that when p_0 data is concentrated in the bottom left corner, the final D still has local maxima outside the support of p_{data}. These local maxima are not suppressed because they were never discovered by the perturbed p_0 data.

The above analysis reveals how binary AT learns data distributions: *the learning starts with a randomly-initialized D solution, and then iteratively refine the solution by suppressing local maxima outside the support of the observed data.* This process is similar to EBMs training where the model distribution's spurious modes are constantly discovered by MCMC sampling and subsequently suppressed in the model update stage. However, unlike the EBMs likelihood objective Eq. (2), the AT objective Eq. (12) cannot properly learn the density function, but can only capture its support. This is corroborated by the 2D experiment where D outputs $\frac{1}{2}$ uniformly on the support of p_{data} (blue points).

Algorithm 1 Binary Adversarial Training

1: **repeat**
2: Draw samples $\{x_i\}_{i=1}^m$ from p_{data}, and samples $\{x_i^0\}_{i=1}^m$ from p_0.
3: Update $\{x_i^0\}_{i=1}^m$ by performing K steps PGD attack Eq. (8) on each sample. Denote the resulting samples as $\{x_i^*\}_{i=1}^m$.
4: Update D by maximizing $\frac{1}{m}\sum_{i=1}^m \log D(x_i) + \frac{1}{m}\sum_{i=1}^m \log(1 - D(x_i^*))$ (single step).
5: **until** D convergences

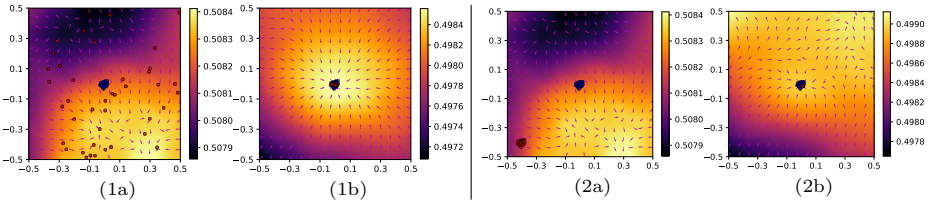

Fig. 1. Plots of contours and (normalized) gradient vector fields of the D functions learned with different p_0 data. Left and right panel respectively show the initial state (1a and 2a) and final state (1b and 2b) of D when p_0 data is respectively uniformly distributed (red points in 1a) and concentrated in the lower left corner (red points in 2a). p_{data} is a Gaussian distribution centered at $(0, 0)$ (blue points).

4.3 Maximum Likelihood Learning Interpretation

We next consider the learning process of binary AT from a maximum likelihood learning point of view. Both binary AT and MCMC-based EBMs learning employ an iterative optimization algorithm, where in each iteration the contrastive data is computed by performing gradient ascent on the current model, and then the model is updated by maximizing its outputs on the observed data and minimizing its outputs on the contrastive data. The following analysis shows that the PGD attack can be viewed as a non-convergent sampler of the model distribution, and the binary AT objective Eq. (11) can be interpreted as a gradient-scaled version of the EBMs objective Eq. (2). Tab. 1 summaries their key differences.

Table 1. Key differences between binary AT and maximum likelihood EBMs.

Objective gradient	EBMs: $\mathbb{E}_{\mathrm{x}\sim p_{\text{data}}}[\nabla_\theta f_\theta(x)] - \mathbb{E}_{\mathrm{x}\sim p_\theta(x)}[\nabla_\theta f_\theta(x)]$
	Binary AT: $\mathbb{E}_{\mathrm{x}\sim p_{\text{data}}}[(1-\sigma(f_\theta(x)))\nabla_\theta f_\theta(x)] - \mathbb{E}_{\mathrm{x}\sim p_T^*}[\sigma(f_\theta(x))\nabla_\theta f_\theta(x)]$
Contrastive data	EBMs: $x_0 \sim p_0, x_{i+1} = x_i + \frac{\lambda}{2}\nabla_x f_\theta(x_i) + \epsilon, \epsilon \sim \mathcal{N}(0, \lambda)$
	Binary AT: $x_0 \sim p_0, x_{i+1} = x_i + \lambda \frac{\nabla_x f_\theta(x_i)}{\|\nabla_x f_\theta(x_i)\|_2}$
p_0 **data**	EBMs: A noise distribution or a distribution close to p_{data}
	Binary AT: A real and diverse out-distribution dataset (80 million tiny images for CIFAR-10 and ImageNet for 256x256 datasets).

Contrastive Data Computation. In EBMs training, the contrastive data is computed by MCMC-sampling, typically with Langevin dynamics Eq. (3). In binary AT, the contrastive data is computed using the PGD attack Eq. (8). Comparing Eq. (8) with Eq. (3), we find that both approaches compute the contrastive data by first initializing from some out-distribution data, and then performing gradient ascent on f_θ. The main differences are that the PGD attack does not have the noise term, and makes use of normalized gradient. Intuitively, the noise term enables the sampler to explore different modes by helping gradient ascent escape local maxima. Although the PGD attack does not have the noise

term, its ability to explore different modes can be enhanced by using a diverse p_0 dataset (Fig. 3).

In the PGD attack, as the normalized gradient vector has unit norm, the perturbation imposed on x_i is λ; in a K iterations of the update, the overall perturbation $\|x_i^* - x_i\|_2$ is always $\leq \lambda K$. Hence with the PGD attack we can more easily control the distribution of the contrastive data. In contrast, Langevin dynamics adjusts x_i in a scale that corresponds to the magnitude of the gradient of f_θ at x_i; when f_θ is updated during training, the overall perturbation may undergo a large change. This behavior of Langevin dynamics can be a source of some training stability issues [39].

Gradient of the Training Objective. By definition Eq. (5), the gradient of D's training objective Eq. (11) takes the form

$$\nabla_\theta J(D) = \mathbb{E}_{\mathrm{x}\sim p_{\mathrm{data}}}[(1-\sigma(f_\theta(x)))\nabla_\theta f_\theta(x)] - \mathbb{E}_{\mathrm{x}\sim p_T^*}[\sigma(f_\theta(x))\nabla_\theta f_\theta(x)]. \quad (14)$$

Comparing the above equation with Eq. (2) we find both equations consisting of gradient terms that yield similar effects: the first term causes f_θ outputs on p_{data} samples to increase, and the second causes f_θ outputs on the contrastive samples to decrease. Specifically, as $(1-\sigma(f_\theta(x)))$ and $\sigma(f_\theta(x))$ are scalars in the range 0 to 1, the two gradient terms in Eq. (14) are respectively the scaled versions of the gradient terms in Eq. (2). It should be noted that although these scalars do not change the gradient update direction of individual terms in the model parameter space, the overall gradient update directions of Eq. (14) and Eq. (2) can be different.

Eq. (14) also helps to understand why binary AT can only learn the support of the observed data. In Eq. (2), when $p_\theta(x)$ matches p_{data}, the gradient cancels out and training terminates, whereas in Eq. (14), when p_T^* matches p_{data} the gradient becomes $\mathbb{E}_{\mathrm{x}\sim p_{\mathrm{data}}}[(1-2\sigma(f_\theta(x)))\nabla_\theta f_\theta(x)]$ and only vanishes when $\sigma(f_\theta(x)) = \frac{1}{2}$ everywhere on the support of p_{data}. This result is consistent with Proposition 1 and the 2D experiment result.

4.4 Improved Training of Binary AT Generative Model

Diverse p_0 Data. As discussed in Sec. 4.2, a diverse p_0 dataset improve the PGD attack's ability to explore different local maxima of D. To validate this 2D intuition generalizes to high dimensions, we evaluate the CIFAR-10 image generation performance under different settings of p_0. It can be seen from Fig. 3 that the best FID is obtained when p_0 is the 80 Million Tiny Images dataset [52], the most diverse dataset among the three p_0 datasets.

We follow existing work on adversarial training and use a p_0 dataset that contains real data samples to train the model. Using a real dataset (as opposed to a noise distribution) helps the model achieve out-of-distribution adversarial robustness (Sec. 5.2 OOD detection) and learn informative gradient for transforming real out-distribution samples (not just noise samples) into valid samples of p_{data}. The latter can be a useful feature in image translation applications (Sec. 5.2). The setting of p_0 in our experiments can be found in Tab. 1.

Training With Unconstrained Perturbations Existing work on using adversarial training to train robust classifiers uses a small, fixed perturbation limit [34]. In the generative modeling task, we would like the perturbed p_0 data to travel in a larger space to find more local maxima. This can be achieved by taking more PGD attack steps (K) in step 3 of Algorithm 1. Fig. 4 shows that a larger K results in better FID scores.

The downside of a large K is that it converges slower because more gradient steps are taken in each iteration (Fig. 5 $K = 25$ vs. $K = 5$). To improve the training efficiency we propose a mixed scenario in which we progressively increase the K value during training. We observe that this progressive training scenario converges faster than training with fixed-K (Fig. 5 $K = 0, 1, ..., 25$ vs $K = 25$). The pseudo code for progressive training is in Algorithm 2.

Algorithm 2 Progressive Binary Adversarial Training

1: **for** K in $[0, 1, \ldots, N]$ **do**
2: **for** number of training iterations **do**
3: Draw samples $\{x_i\}_{i=1}^m$ from p_{data}, and samples $\{x_i^0\}_{i=1}^m$ from p_0.
4: Update $\{x_i^0\}_{i=1}^m$ by performing K steps unconstrained PGD attack Eq. (8) on each sample. Denote the resulting samples as $\{x_i^*\}_{i=1}^m$.
5: Update D by maximizing $\frac{1}{m}\sum_{i=1}^m \log D(x_i) + \frac{1}{m}\sum_{i=1}^m \log(1 - D(x_i^*))$.
6: **end for**
7: **end for**

Failure Mode. When we use Algorithm 1 to train on CelebA-HQ 256 [27], we observe that the binary classification accuracy quickly reaches 100% after a few hundred iterations. Meanwhile the mean l^2 distance between the original p_0 samples and perturbed p_0 samples is only a small fraction of λK, and the perturbed p_0 samples show no meaningful features of human faces. Although the symptoms are different, we believe this failure has the same root cause as mode collapse in GANs. In the adversarial game Eq. (12), the D model can learn to pick up a handful of low-level features which correlate well with the 0-1 labels to solve the binary classification task. With these features the D model cannot provide sufficient gradient for the adversarial p_0 data to move towards the manifold of p_{data}. Without the accurate information, the p_0 adversary fails to compete with D, and the D model starts to dominate the adversarial game. We follow GANs literature [36,4,28] and address this issue by using R_1 regularization [36] to regularize the sensitivity of D's output to the input features (Fig. 2 shows the effect of this regularizer). We do not observe other stability issues other than this failure mode.

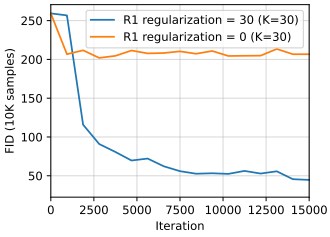

Fig. 2. The effect of R_1 regularization on CelebA-HQ.

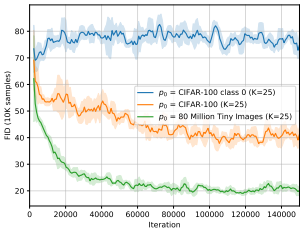

Fig. 3. FID scores obtained with different settings of p_0 on CIFAR-10.

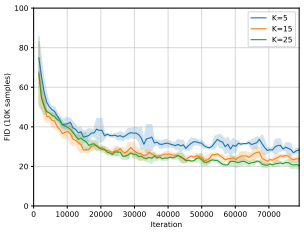

Fig. 4. FID scores obtained with different Ks in Algorithm 1 on CIFAR-10.

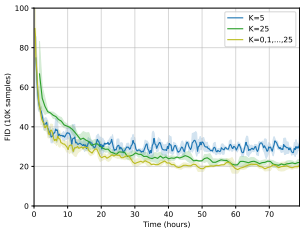

Fig. 5. Progressive training vs training with fixed-K on CIFAR-10.

5 Experiments

In this section we provide an empirical evaluation of the proposed AT generative model. We first evaluate the approach's image generation performance and then demonstrate its applications to image-to-image translation and worst-case out-of-distribution detection. We further provide an analysis of the proposed approach's training stability in Sec. 5.3.

In the supplementary materials we provide experiment setup details including model architectures, training hyperparameters, sample generation settings, and evaluation protocols. We also include additional results including sampling efficiency analysis, uncurated image generation and image translation results, and demonstration of applications to denoising, inpainting, and compositional visual generation [8]. The interpolation results and nearest neighbor analysis in the supplementary materials suggest that our model captures the manifold structure of the observed data, as opposed to simply memorizes the data samples.

5.1 Image Generation

Tab. 2 shows that on CIFAR-10 [30] our approach achieves the best Inception Score (IS) [45] and FID [22] among AT generative models. Our approach also improves over state-of-the-art explicit EBMs in terms of IS, and at the same time has a slightly worse FID. Compared to VAEBM [57], our method does not require an auxiliary model to train, and has better test time sampling efficiency (see supplementary materials). Diffusion Recovery [12] trains a sequence of conditional EBMs, with each one defining the conditional distribution of a noisy sample given the same sample at a higher noise level. Similar to score-based approaches, these conditional EBMs do not directly model the data distribution of the observed data, so it is unclear how these models can be applied to tasks which require explicit knowledge of the data distribution (e.g., OOD detection).

Tab. 3 shows that on CelebA-HQ 256 [27] our method outperforms or is on par with state-of-the-art generative models except GANs. On LSUN Church [66] our method outperforms a latest energy-based model VAEBM [57] (the authors only provided the 64×64 result), but falls bellow DDPM and GANs. Fig. 6 shows sample image generation results.

Table 2. IS and FID scores on CIFAR-10

	Method	IS↑	FID↓
AT	Ours	9.10	13.21
	CEM [55]	8.68	36.4
	Adv. Robust Classifier [47]	7.5	-
Explicit EBMs	Diffusion Recovery [12]	8.30	9.58
	VAEBM [57]	8.43	12.19
	CoopFlow(pretrained Flow) [64]	-	15.80
	CF-EBM [67]	-	16.71
	ImprovedCD [9]	7.85	25.1
	VERA [15]	-	27.5
	EBMs-VAE [63]	6.65	36.2
	JEM [14]	8.76	38.4
	IGEBM (Ensemble) [10]	6.78	38.2
	Short-Run EBMs [40]	6.21	44.16
GANs	StyleGAN2 w/o ADA [28]	8.99	9.9
	BigGAN [4]	9.22	14.73
	SNGAN [37]	8.22	21.7
	WGAN-GP [16]	7.86	36.4
Score-based	SDE [51]	9.89	2.20
	DDPM [23]	9.46	3.17
	NCSNv2 [50]	8.4	10.87
	NCSN [49]	8.87	25.32

Table 3. FID scores on CelebA-HQ 256, AFHQ-CAT [6], and LSUN Church 256

Dataset	Method	FID↓
CelebA-HQ 256	Ours	17.31
	VAEBM [57]	20.38
	CF-EBM [67] (128 × 128)	23.50
	NVAE [54]	45.11
	GLOW [29]	68.93
	Adversarial Latent AE [42]	19.21
	ProgressiveGAN [27]	8.03
AFHQ-CAT	Our (256 × 256)	13.35
	StyleGAN2 (512 × 512) [28]	5.13
LSUN Church	VAEBM (64 × 64) [57]	13.51
	Ours (64 × 64)	10.84
	Ours (256 × 256)	14.87
	DDPM (256 × 256) [23]	7.89
	ProgressiveGAN (256 × 256) [27]	6.42

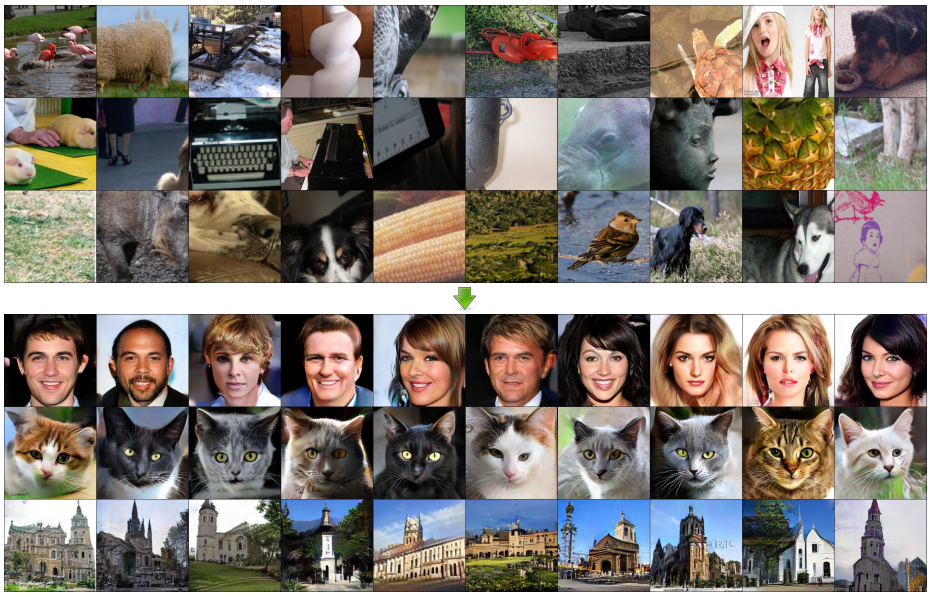

Fig. 6. Source images (top panel) and generated images (bottom panel, 256 × 256 resolution) on CelebA-HQ, AFHQ-CAT, and LSUN Church.

5.2 Applications

Image-To-Image Translation. Fig. 7 shows that the AFHQ-CAT model can be used to transform CelebA-HQ images into cat images, and vice-versa. Note that these two models are trained independently without knowledge of the source

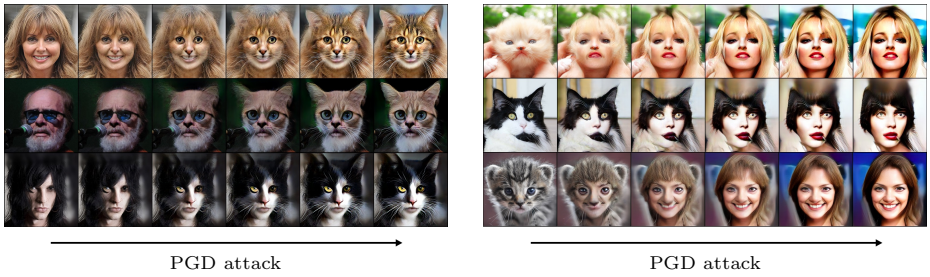

Fig. 7. Image-to-image translation demonstration

domain, indicating that our approach may generalize better to unseen data than approaches (e.g., pix2pix [26], CycleGAN [68], and StarGAN [6]) that explicitly use the source domain dataset to train the model. The translation results may be further improved by finetuning the trained model on the source domain dataset, or including the source domain data in the p_0 dataset during training. The proposed approach is also more flexible than approaches that employs a fixed generator, as it allows the user to choose how much transformation to apply, and/or create cinematic effect from intermediate transformation results.

Worst-Case Out-Of-Distribution Detection. Out-of-distribution (OOD) detection is a classic application of EBMs. Some recent works [48,35,3,2] find that EBMs and some other OOD detection approaches such as OE [21] are vulnerable to adversarial OOD inputs. Given the challenge of adversarial inputs, many authors attempt to address OOD detection in an adversarial setting (also known as worst-case, or adversarial OOD detection) [48,20,35,3,2]. Among these works, RATIO [2] is a state-of-the-art method that combines in- and out-distribution adversarial training to obtain a robust classifier that has uniform outputs in a neighborhood around OOD samples. Similar to OE, RATIO employs 80 million tiny images [52] as the out-distribution dataset to train the model.

Tab. 4 shows that our model achieves comparable OOD detection performance to the state-of-the-art method of RATIO [2]. OE, RATIO, and JEM all perform OOD detection by utilizing a classifier that has low confidence predictions on the out-distribution data (clean or worst-case). In RATIO, the worst-case out-distribution data is computed by performing the PGD attack on 80 million tiny images, whereas in JEM it is computed via Langevin dynamics initialized from uniform random noise. RATIO and our method's strong out-distribution adversarial robustness demonstrates the benefit of using a real and diverse out-distribution dataset to train the model. Our method does not make use of class labels and therefore can be considered as a binary variant of RATIO. On CelebA-HQ 256, AFHQ-CAT, and LSUN-Church our model similarly achieves strong out-distribution adversarial robustness (supplementary materials). These results suggest that our generative model can be applied to detect both naturally occurring OOD data and adversarially created malicious content.

Table 4. CIFAR-10 standard and worst-case OOD detection results (AUC scores). We use the same settings of AutoAttack [7], number of OOD samples, and perturbation limit as in [2] to compute adversarial OOD samples. Results of OE, JEM, and RATIO are from [2].

OOD dataset	Classifier-based approach			Ours
	OE [21]	JEM [14]	RATIO [2]	
Standard OOD				
SVHN	99.4	89.3	96.5	93.0
CIFAR-100	91.4	87.6	91.6	88.3
ImageNet	89.8	86.7	91.3	89.7
Uniform Noise	99.5	11.8	99.9	100
Worst-case OOD				
SVHN	0.6	7.3	81.3	81.6
CIFAR-100	2.7	19.2	73.0	70.3
ImageNet	1.5	21.2	73.5	71.2
Uniform Noise	43.1	2.5	99.8	100

5.3 Training Stability Analysis

To gain some insight into the training stability of our approach we investigate whether the PGD attack can be used with the EBMs training objective Eq. (2). Specifically, in Algorithm 2, we perform step 5's update on θ using the gradient $\nabla_\theta(\frac{1}{m}\sum_{i=1}^m f_\theta(x_i) - \frac{1}{m}\sum_{i=1}^m f_\theta(x_i^*))$. We observe that even under a small learning rate of $1e-6$, $\frac{1}{m}\sum_{i=1}^m f_\theta(x_i) - \frac{1}{m}\sum_{i=1}^m f_\theta(x_i^*)$ quickly increases and eventually overflows. This suggests that the stability of the AT approach can be largely attributed to the log-likelihood objective Eq. (11). We argue that the stability is due to the gradient cancelling effect of this objective: when f_θ has a large positive output on a sample $x \sim p_{\text{data}}$, $1 - \sigma(f_\theta(x))$ approaches 0 and therefore the corresponding scaled gradient in Eq. (14) vanishes, and similarly $\sigma(f_\theta(x^*))\nabla_\theta f_\theta(x^*)$ vanishes when f_θ has a large negative output on a sample $x^* \sim p_T^*$. In contrast, the EBMs objective Eq. (2) does not have constraints on f_θ's outputs and is therefore prone to divergence.

6 Conclusion

We have studied an AT-based approach to learning EBMs. Our analysis shows that binary AT learns a special kind of energy function that models the support of the observed data, and the training procedure can be viewed as an approximate maximum likelihood learning algorithm. We proposed improved techniques for generative modeling with AT, and demonstrated that the proposed method provides competitive generation performance to explicit EBMs, has competitive sampling efficiency, is stable to train, and is well-suited for image translation tasks. The proposed approach's strong out-distribution adversarial robustness suggests its potential application to detecting abnormal inputs and/or adversarially created fake content.

References

1. Arbel, M., Zhou, L., Gretton, A.: Generalized energy based models. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=0PtUPB9z6qK 3
2. Augustin, M., Meinke, A., Hein, M.: Adversarial robustness on in-and out-distribution improves explainability. In: European Conference on Computer Vision. pp. 228–245. Springer (2020) 13, 14
3. Bitterwolf, J., Meinke, A., Hein, M.: Certifiably adversarially robust detection of out-of-distribution data. Advances in Neural Information Processing Systems **33** (2020) 13
4. Brock, A., Donahue, J., Simonyan, K.: Large scale GAN training for high fidelity natural image synthesis. In: International Conference on Learning Representations (2019), https://openreview.net/forum?id=B1xsqj09Fm 10, 12
5. Ceylan, C., Gutmann, M.U.: Conditional noise-contrastive estimation of unnormalised models. In: International Conference on Machine Learning. pp. 726–734. PMLR (2018) 4
6. Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: Stargan v2: Diverse image synthesis for multiple domains. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8188–8197 (2020) 12, 13
7. Croce, F., Hein, M.: Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: ICML (2020) 14
8. Du, Y., Li, S., Mordatch, I.: Compositional visual generation with energy based models. Advances in Neural Information Processing Systems **33**, 6637–6647 (2020) 11
9. Du, Y., Li, S., Tenenbaum, J.B., Mordatch, I.: Improved contrastive divergence training of energy based models. In: ICML (2021) 1, 3, 12
10. Du, Y., Mordatch, I.: Implicit generation and modeling with energy based models. In: Advances in Neural Information Processing Systems. vol. 32 (2019), https://proceedings.neurips.cc/paper/2019/file/378a063b8fdb1db941e34f4bde584c7d-Paper.pdf 1, 3, 4, 12
11. Engstrom, L., Ilyas, A., Santurkar, S., Tsipras, D., Tran, B., Madry, A.: Adversarial robustness as a prior for learned representations. arXiv preprint arXiv:1906.00945 (2019) 2
12. Gao, R., Song, Y., Poole, B., Wu, Y.N., Kingma, D.P.: Learning energy-based models by diffusion recovery likelihood. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=v_1Soh8QUNc 11, 12
13. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014) 6
14. Grathwohl, W., Wang, K.C., Jacobsen, J.H., Duvenaud, D., Norouzi, M., Swersky, K.: Your classifier is secretly an energy based model and you should treat it like one. In: International Conference on Learning Representations (2020), https://openreview.net/forum?id=Hkxzx0NtDB 1, 3, 4, 12, 14
15. Grathwohl, W.S., Kelly, J.J., Hashemi, M., Norouzi, M., Swersky, K., Duvenaud, D.: No mcmc for me: Amortized sampling for fast and stable training of energy-based models. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=ixpSxO9flk3 1, 3, 12

16. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. In: Advances in Neural Information Processing Systems. vol. 30 (2017), https://proceedings.neurips.cc/paper/2017/file/892c3b1c6dccd52936e27cbd0ff683d6-Paper.pdf 12
17. Gutmann, M., Hyvärinen, A.: Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics. pp. 297–304. JMLR Workshop and Conference Proceedings (2010) 3, 4
18. Han, T., Nijkamp, E., Fang, X., Hill, M., Zhu, S.C., Wu, Y.N.: Divergence triangle for joint training of generator model, energy-based model, and inferential model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8670–8679 (2019) 3
19. Han, T., Nijkamp, E., Zhou, L., Pang, B., Zhu, S.C., Wu, Y.N.: Joint training of variational auto-encoder and latent energy-based model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7978–7987 (2020) 3
20. Hein, M., Andriushchenko, M., Bitterwolf, J.: Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 41–50 (2019) 13
21. Hendrycks, D., Mazeika, M., Dietterich, T.: Deep anomaly detection with outlier exposure. arXiv preprint arXiv:1812.04606 (2018) 13, 14
22. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in neural information processing systems. pp. 6626–6637 (2017) 11
23. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems. vol. 33 (2020), https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf 3, 12
24. Hyvärinen, A., Dayan, P.: Estimation of non-normalized statistical models by score matching. Journal of Machine Learning Research **6**(4) (2005) 3
25. Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., Madry, A.: Adversarial examples are not bugs, they are features. arXiv preprint arXiv:1905.02175 (2019) 2
26. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017) 13
27. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017) 10, 11, 12
28. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. In: Advances in Neural Information Processing Systems. vol. 33, pp. 12104–12114 (2020), https://proceedings.neurips.cc/paper/2020/file/8d30aa96e72440759f74bd2306c1fa3d-Paper.pdf 10, 12
29. Kingma, D.P., Dhariwal, P.: Glow: Generative flow with invertible 1x1 convolutions. In: Advances in neural information processing systems. pp. 10215–10224 (2018) 12
30. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009) 11

31. Kumar, R., Ozair, S., Goyal, A., Courville, A., Bengio, Y.: Maximum entropy generators for energy-based models. arXiv preprint arXiv:1901.08508 (2019) 3
32. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236 (2016) 5
33. LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., Huang, F.: A tutorial on energy-based learning. Predicting structured data **1**(0) (2006) 1, 4
34. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017) 5, 10
35. Meinke, A., Hein, M.: Towards neural networks that provably know when they don't know. arXiv preprint arXiv:1909.12180 (2019) 13
36. Mescheder, L., Geiger, A., Nowozin, S.: Which training methods for gans do actually converge? In: International conference on machine learning. pp. 3481–3490. PMLR (2018) 10
37. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. In: International Conference on Learning Representations (2018), https://openreview.net/forum?id=B1QRgziT- 12
38. Nijkamp, E., Gao, R., Sountsov, P., Vasudevan, S., Pang, B., Zhu, S.C., Wu, Y.N.: MCMC should mix: Learning energy-based model with flow-based backbone. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=4C93Qvn-tz 3
39. Nijkamp, E., Hill, M., Han, T., Zhu, S.C., Wu, Y.N.: On the anatomy of mcmc-based maximum likelihood learning of energy-based models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 5272–5280 (2020) 1, 2, 3, 4, 9
40. Nijkamp, E., Hill, M., Zhu, S.C., Wu, Y.N.: Learning non-convergent non-persistent short-run mcmc toward energy-based model. In: NeurIPS (2019) 1, 2, 3, 4, 12
41. Pang, B., Han, T., Nijkamp, E., Zhu, S.C., Wu, Y.N.: Learning latent space energy-based prior model. In: Advances in Neural Information Processing Systems. vol. 33 (2020), https://proceedings.neurips.cc/paper/2020/file/fa3060edb66e6ff4507886f9912e1ab9-Paper.pdf 3
42. Pidhorskyi, S., Adjeroh, D.A., Doretto, G.: Adversarial latent autoencoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14104–14113 (2020) 12
43. Ramachandran, P., Zoph, B., Le, Q.V.: Swish: a self-gated activation function. arXiv preprint arXiv:1710.05941 **7**, 1 (2017) 3
44. Rhodes, B., Xu, K., Gutmann, M.U.: Telescoping density-ratio estimation. arXiv preprint arXiv:2006.12204 (2020) 4
45. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. Advances in neural information processing systems **29**, 2234–2242 (2016) 11
46. Salimans, T., Kingma, D.P.: Weight normalization: A simple reparameterization to accelerate training of deep neural networks. Advances in neural information processing systems **29**, 901–909 (2016) 3
47. Santurkar, S., Ilyas, A., Tsipras, D., Engstrom, L., Tran, B., Madry, A.: Image synthesis with a single (robust) classifier. In: Advances in Neural Information Processing Systems. pp. 1260–1271 (2019) 2, 12
48. Sehwag, V., Bhagoji, A.N., Song, L., Sitawarin, C., Cullina, D., Chiang, M., Mittal, P.: Better the devil you know: An analysis of evasion attacks using out-of-distribution adversarial examples. arXiv preprint arXiv:1905.01726 (2019) 13

49. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. In: Advances in Neural Information Processing Systems. vol. 32 (2019), https://proceedings.neurips.cc/paper/2019/file/3001ef257407d5a371a96dcd947c7d93-Paper.pdf 3, 12
50. Song, Y., Ermon, S.: Improved techniques for training score-based generative models. arXiv preprint arXiv:2006.09011 (2020) 3, 12
51. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=PxTIG12RRHS 3, 12
52. Torralba, A., Fergus, R., Freeman, W.T.: 80 million tiny images: A large data set for nonparametric object and scene recognition. IEEE transactions on pattern analysis and machine intelligence **30**(11), 1958–1970 (2008) 9, 13
53. Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., Madry, A.: Robustness may be at odds with accuracy. arXiv preprint arXiv:1805.12152 (2018) 2
54. Vahdat, A., Kautz, J.: Nvae: A deep hierarchical variational autoencoder. In: Advances in Neural Information Processing Systems. vol. 33 (2020), https://proceedings.neurips.cc/paper/2020/file/e3b21256183cf7c2c7a66be163579d37-Paper.pdf 12
55. Wang, Y., Wang, Y., Yang, J., Lin, Z.: A unified contrastive energy-based model for understanding the generative ability of adversarial training. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=XhF2VOMRHS 12
56. Welling, M., Teh, Y.W.: Bayesian learning via stochastic gradient langevin dynamics. In: Proceedings of the 28th international conference on machine learning (ICML-11). pp. 681–688. Citeseer (2011) 3, 4
57. Xiao, Z., Kreis, K., Kautz, J., Vahdat, A.: Vaebm: A symbiosis between variational autoencoders and energy-based models. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=5m3SEczOV8L 1, 3, 11, 12
58. Xie, J., Lu, Y., Gao, R., Zhu, S.C., Wu, Y.N.: Cooperative training of descriptor and generator networks. IEEE transactions on pattern analysis and machine intelligence **42**(1), 27–45 (2018) 3
59. Xie, J., Lu, Y., Zhu, S.C., Wu, Y.: A theory of generative convnet. In: International Conference on Machine Learning. pp. 2635–2644. PMLR (2016) 3, 4
60. Xie, J., Zheng, Z., Fang, X., Zhu, S.C., Wu, Y.N.: Cooperative training of fast thinking initializer and slow thinking solver for conditional learning. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021) 4
61. Xie, J., Zheng, Z., Gao, R., Wang, W., Zhu, S.C., Wu, Y.N.: Learning descriptor networks for 3d shape synthesis and analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8629–8638 (2018) 4
62. Xie, J., Zheng, Z., Gao, R., Wang, W., Zhu, S.C., Wu, Y.N.: Generative voxelnet: learning energy-based models for 3d shape synthesis and analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020) 4
63. Xie, J., Zheng, Z., Li, P.: Learning energybased model with variational auto-encoder as amortized sampler. In: The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI). vol. 2 (2021) 3, 12
64. Xie, J., Zhu, Y., Li, J., Li, P.: A tale of two flows: Cooperative learning of langevin flow and normalizing flow toward energy-based model. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=31d5RLCUuXC 3, 12

65. Yin, X., Kolouri, S., Rohde, G.K.: Gat: Generative adversarial training for adversarial example detection and robust classification. In: International Conference on Learning Representations (2020), https://openreview.net/forum?id=SJeQEp4YDH 2, 5
66. Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., Xiao, J.: Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365 (2015) 11
67. Zhao, Y., Xie, J., Li, P.: Learning energy-based generative models via coarse-to-fine expanding and sampling. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=aD1_5zowqV 1, 3, 12
68. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017) 13