

EGC: Image Generation and Classification via a Diffusion Energy-Based Model

Qiushan Guo¹, Chuofan Ma¹, Yi Jiang², Zehuan Yuan², Yizhou Yu¹, Ping Luo^{1,3}

¹The University of Hong Kong ²ByteDance Inc. ³Shanghai AI Laboratory



Figure 1: Generated samples from our EGC models on CelebA-HQ 1024×1024, LSUN Church 256×256, CelebA-HQ 256×256 and ImageNet 256×256 datasets, shown from left to right.

Abstract

Learning image classification and image generation using the same set of network parameters presents a formidable challenge. Recent advanced approaches perform well in one task often exhibit poor performance in the other. This work introduces an energy-based classifier and generator, namely EGC, which can achieve superior performance in both tasks using a single neural network. Unlike conventional classifiers that produce a label given an image (i.e., a conditional distribution $p(y|x)$), the forward pass in EGC is a classification model that yields a joint distribution $p(x, y)$, enabling a diffusion model in its backward pass by marginalizing out the label y to estimate the score function. Furthermore, EGC can be adapted for unsupervised learning by considering the label as latent variables. EGC achieves competitive generation results compared with state-of-the-art approaches on ImageNet-1k, CelebA-HQ and LSUN Church, while achieving superior classification accuracy and robustness against adversarial attacks on CIFAR-10. This work marks the inaugural success in mastering both domains using a unified network parameter set. We believe that EGC bridges the gap between discriminative and generative learning. Code will be released at <https://github.com/GuoQiushan/EGC>.

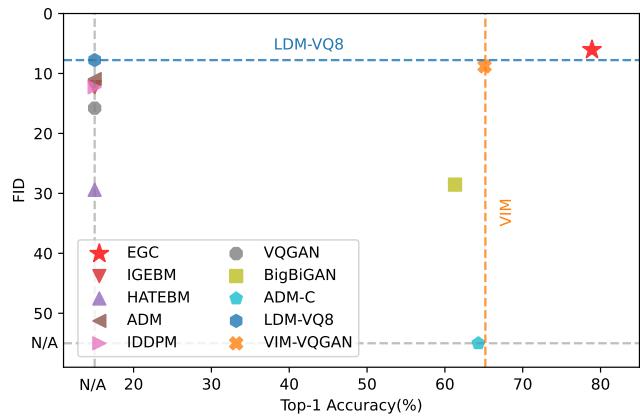


Figure 2: FID and classification accuracy on ImageNet 256×256 dataset. The scatters plots on the vertical N/A line represent the image generation models, which are not available for classification. The scatter plot on the horizontal N/A line represents the classification model, which is not available for image generation. Remarkably, EGC achieves superior performance in both tasks with a single neural network, demonstrating its effectiveness in bridging the gap between discriminative and generative learning.

1. Introduction

Image classification and generation are two fundamental tasks in computer vision that have seen significant advancements with the development of deep learning models. However, many state-of-the-art approaches that perform well in one task often exhibit poor performance in the other or are not suitable for the other task. Both tasks can be formulated from a probabilistic perspective, where image classification task is interpreted as a conditional probability distribution $p(y|\mathbf{x})$, and image generation task is the transformation of a known and easy-to-sample probability distribution $p(\mathbf{z})$ to a target distribution $p(\mathbf{x})$.

As an appealing class of probabilistic models, Energy-Based Models (EBM) [6, 9, 10, 11, 12, 15, 16, 29, 32, 41, 43, 55, 60, 65, 68] can explicitly model complex probability distribution. Furthermore, standard image classification models and some image generation models can be reinterpreted as EBMs [12, 16, 24, 33, 36, 37]. JEM [16] repurposes a standard image classification model as an image generation model by leveraging the gradient of its inputs to guide the generation of new images. Despite the desirable properties, EBMs face challenges in training due to the intractability of computing the exact likelihood and synthesizing exact samples from these models. Arbitrary energy models often exhibit sharp changes in gradients, leading to unstable sampling with Langevin dynamics. To ameliorate this issue, spectral normalization [40] is typically adopted for constraining the Lipschitz constant of the energy model [6, 12, 16]. Even with this regularization technique, the samples generated by the energy-based model are still not competitive enough because the probability distribution of real data is usually sharp in the high-dimensional space, providing inaccurate guidance for image sampling in the low data density regions.

Diffusion models [21, 45, 49, 50, 51] have demonstrated competitive and even superior image generation performance compared to GAN [13] models. In diffusion models, images are perturbed with Gaussian noise through a diffusion process for training, and the reverse process is learned to transform the Gaussian distribution back to the data distribution. As pointed out in [50], perturbing data points with noise populates low data density regions to improve the accuracy of estimated scores, resulting in stable training and image sampling.

Motivated by the flexibility of JEM and the stability of diffusion model, we propose a novel energy-based classifier and generator, namely EGC, which achieves superior performance in both image classification and generation tasks using a single neural network. EGC is a classifier in the forward pass and an image generator in the backward pass. Unlike a conventional classifier that predicts the condition distribution $p(y|\mathbf{x})$ of the label given an image, the forward pass in EGC models the joint distribution

$p(\mathbf{x}, y)$ of the noisy image and label, given the source image. By marginalizing out the label y , the gradient of log-probability of the noisy image (*i.e.*, unconditional score) is used to restore image from noise. The classification probability $p(y|\mathbf{x})$ provides classifier guidance together with unconditional score within one step backward pass.

We demonstrate the efficacy of EGC model on ImageNet, CIFAR-10, CIFAR-100, CelebA-HQ and LSUN datasets. The generated samples are of high fidelity and comparable to GAN-based methods, as shown in Fig. 1. Additionally, our model shows superior classification accuracy and robustness against adversarial attacks. On CIFAR-10, EGC surpasses existing methods of learning explicit EBMs with an FID of 3.30 and an inception score of 9.43 while achieving a remarkable classification accuracy of 95.9%. This result even exceeds the classification performance of the discriminative model Wide ResNet-28-12, which shares a comparable architecture and number of parameters with our model. On ImageNet-1k, EGC achieves an FID of **6.05** and an accuracy of **78.9%**, as illustrated in Fig. 2. We also demonstrate that naively optimizing the gradients of explicit energy functions as the score functions outperforms optimizing the probability density function via Langevin sampling. Besides, EGC model does not require constraining the Lipschitz constant as in the previous methods [6, 12, 16] by removing the normalization layers and inserting spectral normalization layer. More interestingly, we demonstrate that the neural network effectively models the target data distribution even though we adopt optimization of the Fisher divergence instead of the probability $p_\theta(\mathbf{x}_t)$.

Our contributions are listed as follows:

- (1) We propose a novel hybrid energy-based model, EGC, which is trained in diffusion manner using Fisher divergence. In EGC, the forward pass is a classification model and the backward pass is a diffusion model that denoises data using the score function and conditional guidance.
- (2) Our EGC model achieves competitive generation results to state-of-the-art approaches, while obtaining superior classification results using a single neural network. EGC surpasses existing methods of explicit EBMs by a significant margin.
- (3) We demonstrate that EGC model can be applied in inpainting, semantic interpolation, high-resolution image generation ($\sim 1024^2$) and robustness improvement.

2. Related Works

Energy-based Models. Unlike most other probabilistic models, Energy-Based Models do not place a restriction on the tractability of the normalizing constant, which confers upon them greater flexibility in modeling complex probability distributions. However, the intractability of the normalization constant renders their training challenging. BiDVL [25] proposes a bi-level optimization framework to facilitate learning of energy-based latent variable models. Yin

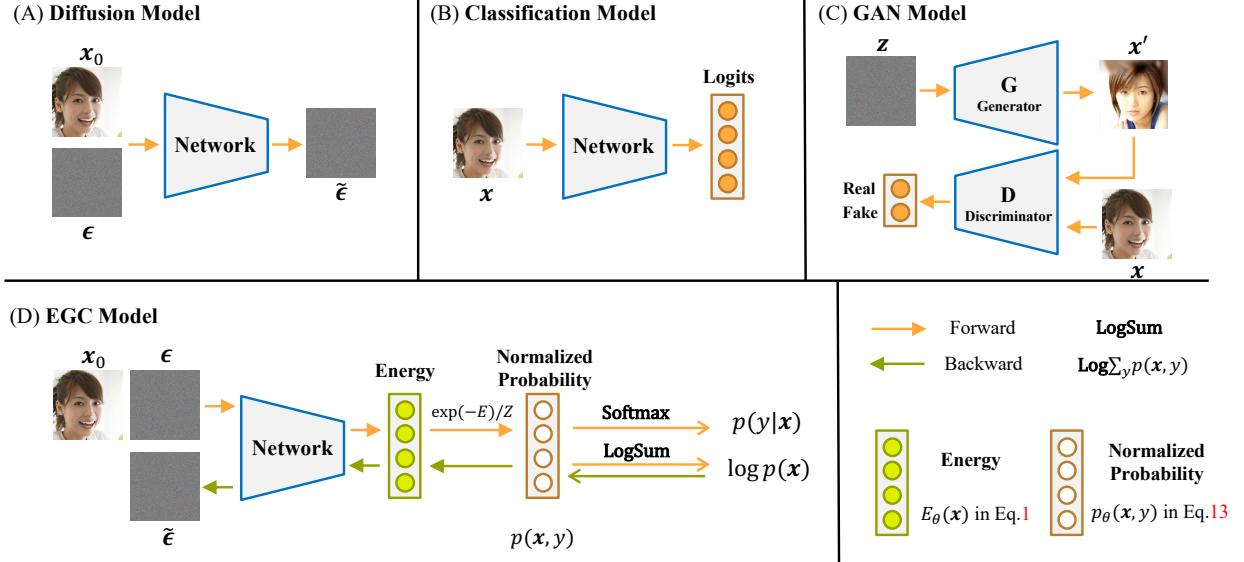


Figure 3: (A) Diffusion Model estimates the score (noise) from the noisy scaled image. (B) Standard Classification Model outputs the logits for minimizing the cross-entropy loss. (C) GAN Model is composed of a generator model that synthesizes new samples and a discriminator that classifies samples as either real or fake. (D) EGC Model estimates the joint distribution $p(\mathbf{x}, \mathbf{y})$ for classification via the forward propagation of a neural network and leverages the score estimated from the backward propagation to generate samples from Gaussian noise. Z represents the normalizing constant, which is only relevant to the model parameters.

et al. [65] explore adversarial training for learning EBMs. CLEL [35] improves the training of EBMs using contrastive representation learning, which guides EBMs to better understand the data structure for faster and more memory-efficient training. [58, 59, 61] are a series of work to perform efficient EBM generation via cooperative learning. In contrast to the above methods, we reinterpret the standard classification network as an EBM to approximate the joint distribution of the noisy samples in diffusion process. Our method enables the forward pass to work as a classification model and the backward pass to provide both unconditional and conditional scores within one step.

Denoising Diffusion Models (DDMs), originating from [48], are to learn from noisy data and generate samples by reversing the diffusion process. Score based generative model [50] is introduced to train denoising models with multiple noise levels and draw samples via Langevin dynamics during inference. Several works have improved the design and demonstrated the capability of synthesizing high-quality images [21, 4, 22]. DDPM[21] generates high-quality images synthesis results using diffusion probabilistic models. Guided-Diffusion [4] firstly achieved better performance than GAN by utilizing the classifier guidance. Latent Diffusion [45] further proposes latent diffusion models which operate on a compressed latent space of reduced dimensionality to save computational cost. Collaborative Diffusion [23] follows the collaborative learning [17] to achieve multi-modal face generation and editing. DDMs

circumvent the issue of intractable normalizing constants in EBMs by modeling the score function instead of the density function. Different from the score-based diffusion models, our EGC explicitly models the probability distribution, and both forward and backward passes are meaningful.

Unified Classification and Generation Models. Recent advanced approaches [18, 46, 64] perform well in one task often exhibit poor performance in the other. Xie *et al.* [60] first draw the connection between the discriminative and generative power of a ConveNet with EBMs. Based on the idea, Grathwohl *et al.* [16] propose to re-interpret a discriminative classifier as an EBM modeling joint distribution of \mathbf{x} and \mathbf{y} , which enables image synthesis and classification within one framework. Introspective Neural Networks [24, 33, 36] share a similar insight to imbue the classifier with generative power, leading to increased robustness in adversarial attacks. In contrast to the above methods, our method adopts the diffusion process to improve the accuracy of estimated scores for stable training and image sampling. Consequently, the regularization and training tricks required by other models are not necessary for our EGC model. EGC significantly outperforms the hybrid model, JEM [16], by a substantial margin.

3. Method

Overview. As illustrated in Fig. 3, the EGC model consists of a classifier that models the joint distribution $p(\mathbf{x}, \mathbf{y})$

by estimating the energy during the forward pass. The conditional probability $p(y|\mathbf{x})$ is produced by Softmax function. The backward pass of the network produces both the unconditional score ($\nabla_{\mathbf{x}} \log p(\mathbf{x})$) and the class guidance ($\nabla_{\mathbf{x}} \log p(y|\mathbf{x})$) in a single step by marginalizing out y . We adopt the diffusion process to populate low data density regions and optimize the unconditional score using Fisher divergence, which circumvents the direct optimization of normalized probability density.

Background. An energy-based model [34] is defined as

$$p_{\theta}(\mathbf{x}) = \frac{\exp(-E_{\theta}(\mathbf{x}))}{Z(\theta)}, \quad (1)$$

to approximate any probability density function $p_{\text{data}}(\mathbf{x})$ for $\mathbf{x} \in \mathbb{R}^D$, where $E_{\theta}(\mathbf{x}) : \mathbb{R}^D \rightarrow \mathbb{R}$, known as the energy function, maps each data point to a scalar, and $Z(\theta) = \int \exp(-E_{\theta}(\mathbf{x})) d\mathbf{x}$, analytically intractable for high-dimensional \mathbf{x} , is the partition function. Typically, one can parameterize the energy function with a neural network $f_{\theta}(\mathbf{x}) = -E_{\theta}(\mathbf{x})$. The *de facto* standard for learning probabilistic models from *i.i.d* data is maximum likelihood estimation. The log-likelihood function is

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [\log p_{\theta}(\mathbf{x})] \simeq \frac{1}{N} \sum_{i=1}^N \log p_{\theta}(\mathbf{x}_i), \quad (2)$$

where we observe N samples $\mathbf{x}_i \sim p_{\text{data}}(\mathbf{x})$. The gradient of the log-probability of an EBM is composed of two terms:

$$\frac{\partial \log p_{\theta}(\mathbf{x})}{\partial \theta} = \frac{\partial f_{\theta}(\mathbf{x})}{\partial \theta} - \mathbb{E}_{\mathbf{x}' \sim p_{\theta}(\mathbf{x}')} \left[\frac{\partial f_{\theta}(\mathbf{x}')}{\partial \theta} \right], \quad (3)$$

The second term can be approximated by drawing the synthesized samples from the model distribution $p_{\theta}(\mathbf{x}')$ with Markov Chain Monte Carlo (MCMC). Langevin MCMC first draws an initial sample from a simple prior distribution and iteratively updates the sample until a mode is reached, which can be formalized as follows:

$$\mathbf{x}_{i+1} \leftarrow \mathbf{x}_i + c \nabla_{\mathbf{x}} \log p(\mathbf{x}_i) + \sqrt{2c} \boldsymbol{\epsilon}_i, \quad (4)$$

where \mathbf{x}_0 is randomly sampled from a prior distribution (such as Gaussian distribution), and $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. However, for high-dimensional distributions, it takes a long time to run MCMC to generate a converged sample.

3.1. Energy-Based Model with Diffusion Process

Diffusion models gradually inject noise into the source samples during the nosing process, which can be formulated as a Markov chain:

$$\begin{aligned} q(\mathbf{x}_{1:T} | \mathbf{x}_0) &= \prod_{t=1}^T q(\mathbf{x}_t | \mathbf{x}_{t-1}), \\ q(\mathbf{x}_t | \mathbf{x}_{t-1}) &= \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I}), \end{aligned} \quad (5)$$

where \mathbf{x}_0 represents the source samples and $\alpha_t = 1 - \beta_t$. For an arbitrary timestep t , one can directly sample from the following Gaussian distribution without iterative sampling,

$$q(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t} \mathbf{x}_0, (1 - \alpha_t) \mathbf{I}) \quad (6)$$

Based on Bayes theorem, one can reverse the noising process by the posterior $q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0)$:

$$\begin{aligned} \tilde{\beta}_t &= \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t \\ \tilde{\mu}(\mathbf{x}_t, \mathbf{x}_0) &= \frac{\sqrt{\bar{\alpha}_{t-1}} \beta_t}{1 - \bar{\alpha}_t} \mathbf{x}_0 + \frac{\sqrt{\bar{\alpha}_t} (1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} \mathbf{x}_t \end{aligned} \quad (7)$$

$$q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\mu}(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t \mathbf{I})$$

\mathbf{x}_0 is unknown during the denoising process, so we approximate the posterior with $p_{\theta}(\mathbf{x}_{t-1} | \mathbf{x}_t) = q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0 = \mu_{\theta}(\mathbf{x}_t))$ to denoise the observed sample \mathbf{x}_t .

According to Tweedie's Formula [7], one can estimate the mean of a Gaussian distribution, given a random variable $\mathbf{z} \sim \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_z, \Sigma_z)$:

$$\mathbb{E}[\boldsymbol{\mu}_z | \mathbf{z}] = \mathbf{z} + \Sigma_z \nabla_{\mathbf{z}} \log p(\mathbf{z}) \quad (8)$$

By applying Tweedie's Formula to Equation 6, the estimate for the mean of the noised sample \mathbf{x}_t can be represented as:

$$\sqrt{\bar{\alpha}_t} \mathbf{x}_0 = \mathbf{x}_t + (1 - \bar{\alpha}_t) \nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t | \mathbf{x}_0) \quad (9)$$

Since the noised sample \mathbf{x}_t decomposes as a sum of two terms: $\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}_t$, the score function $\nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t | \mathbf{x}_0)$ can be expressed as:

$$\nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t | \mathbf{x}_0) = -\frac{\boldsymbol{\epsilon}_t}{\sqrt{1 - \bar{\alpha}_t}} \quad (10)$$

We approximate the probability density function $q(\mathbf{x}_t | \mathbf{x}_0)$ by an Energy-based model $p_{\theta}(\mathbf{x}_t)$, and optimize the parameters by minimizing the Fisher divergence between $q(\mathbf{x}_t | \mathbf{x}_0)$ and $p_{\theta}(\mathbf{x}_t)$:

$$\mathcal{D}_F = \mathbb{E}_q \left[\frac{1}{2} \|\nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t | \mathbf{x}_0) - \nabla_{\mathbf{x}_t} \log p_{\theta}(\mathbf{x}_t)\|^2 \right] \quad (11)$$

For Energy-based models, the score can be easily obtained, $\nabla_{\mathbf{x}} \log p_{\theta}(\mathbf{x}) = \nabla_{\mathbf{x}} f_{\theta}(\mathbf{x})$. Compared with directly optimizing the log-probability of EBM (Equation 3), Fisher divergence circumvents optimizing the normalized densities parameterized by $Z(\theta)$ and the target score can be directly sampled from a Gaussian distribution via Equation 10.

3.2. EGC

The energy-based model has an inherent connection with discriminative models. For the classification problem with C classes, a discriminative neural classifier maps a data sample $\mathbf{x} \in \mathbb{R}^D$ to a vector of length C known as logits.

The probability of y -th label is represented using the Softmax function:

$$p(y|\mathbf{x}) = \frac{\exp(f(\mathbf{x})[y])}{\sum_{y'} \exp(f(\mathbf{x})[y'])}, \quad (12)$$

where $f(\mathbf{x})[y]$ is the y -th logit. Using Bayes theorem, the discriminative conditional probability can be expressed as $p(y|\mathbf{x}) = \frac{p(\mathbf{x}, y)}{\sum_{y'} p(\mathbf{x}, y')}$. By connecting Equation 1 and 12, the joint probability of data sample \mathbf{x} and label y can be modeled as:

$$p_\theta(\mathbf{x}, y) = \frac{\exp(f_\theta(\mathbf{x})[y])}{Z(\theta)} \quad (13)$$

By marginalizing out y , the score of data sample \mathbf{x} is obtained as:

$$\nabla_{\mathbf{x}} \log p_\theta(\mathbf{x}) = \nabla_{\mathbf{x}} \log \sum_y \exp(f_\theta(\mathbf{x})[y]) \quad (14)$$

Different from the typical classifiers, the free energy function $E_\theta(\mathbf{x}) = -\log \sum_y \exp(f_\theta(\mathbf{x})[y])$ is also optimized for generating samples.

We propose to integrate the energy-based classifier with the diffusion process to achieve both strong discriminative performance and generative performance. Specifically, we approximate the conditional probability density function $q(\mathbf{x}_t, y|\mathbf{x}_0)$ with an energy-based classifier $p_\theta(\mathbf{x}_t, y)$. Due to the optimization of Fisher divergence for our EBM, we factorize the log-likelihood as:

$$\log p_\theta(\mathbf{x}_t, y) = \log p_\theta(\mathbf{x}_t) + \log p_\theta(y|\mathbf{x}_t) \quad (15)$$

The score $\nabla_{\mathbf{x}_t} \log p_\theta(\mathbf{x}_t)$ is optimized by minimizing the Fisher divergence as shown in the Equation 11 and 14. As for the conditional probability $p_\theta(y|\mathbf{x}_t)$, we simply adopt the standard cross-entropy loss to optimize.

One of the advantage of integrating the energy-based classifier with diffusion process is that the classifier provides guidance to explicitly control the data we generate through conditioning information y . By Bayes theorem, the conditional score can be derived as:

$$\nabla \log p_\theta(\mathbf{x}_t|y) = \nabla \log p_\theta(\mathbf{x}_t) + \nabla \log p_\theta(y|\mathbf{x}_t) \quad (16)$$

The joint probability $p_\theta(\mathbf{x}_t, y)$ is parameterized with a neural network. And the forward propagation of our EGC model is a discrimination model to predict the conditional probability $p_\theta(y|\mathbf{x})$, while the backward propagation of the neural network is a generation model to predict the score and classifier guidance to gradually denoise data.

Overall, the training loss of an EGC model is formulated as:

$$\begin{aligned} \mathcal{L} = & \mathbb{E}_q [\frac{1}{2} \|\nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t|\mathbf{x}_0) - \nabla_{\mathbf{x}_t} \log p_\theta(\mathbf{x}_t)\|^2 \\ & - \sum_{i=1}^C q(y_i|\mathbf{x}_t, \mathbf{x}_0) \log p_\theta(y_i|\mathbf{x}_t)], \end{aligned} \quad (17)$$

where the first term is reconstruction loss for a noised sample, the second term is a classification loss that encourages the denoising process to generate samples that are consistent with the given labels. The training procedure is summarized in Algorithm 1, where we adopt the noise ϵ as the target score to ensure the stable optimization of the neural network. Additionally, a hyperparameter γ is introduced to balance the two loss terms. The training of EGC requires the second partial derivatives.

Algorithm 1 Training

repeat

 Sample $t \sim \text{Unif}(\{1, \dots, T\})$
 Sample data pair (\mathbf{x}_0, y) , Sample noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
 $\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$
 Take gradient descent step on $\nabla_\theta (\|\nabla_{\mathbf{x}_t} \log p_\theta(\mathbf{x}_t) + \epsilon\|^2 - \gamma \sum_{i=1}^C q(y_i|\mathbf{x}_t) \log p_\theta(y_i|\mathbf{x}_t))$
until converged.

4. Experiments

We conduct a series of experiments to evaluate the performance of EGC model on image classification and generation benchmarks. The results show that our model achieves performance rivaling the state of the art in both discriminative and generative modeling.

4.1. Experimental Setup

For conditional learning, we consider CIFAR-10 [31], CIFAR-100 and ImageNet [3] dataset to evaluate the proposed EGC model. Both CIFAR-10 and CIFAR-100 contain 50K training images. The ImageNet training set is composed of about 1.28 million images from 1000 different categories. For unconditional learning, we train unsupervised EGC models on CelebA-HQ [26], which contains 30K training images of human faces, and LSUN Church [66], which contains about 125K images of outdoor churches. For ImageNet-1k, CelebA-HQ and LSUN Church, we follow latent diffusion models (LDM) [45] to convert images at 256×256 resolution to latent representations at 32×32 or 64×64 resolutions, using the pre-trained image autoencoder provided by LDM [45]. We adopt the same UNet architecture as [4] and attach an attention pooling module to it, like CLIP [44], to predict logits. More training details can be found in Appendix.

4.2. Hybrid Modeling

EGC model. We first train an EGC model on CIFAR-10 dataset. The accuracy on CIFAR-10 validation dataset, Inception Score (IS) [47] and Frechet Inception Distance (FID) [19] are reported to quantify the performance of our model. Our model demonstrates a remarkable accuracy of 95.9% on the CIFAR-10 validation dataset, surpassing

Method	Acc(%)	IS(\uparrow)	FID(\downarrow)
<i>Hybrid Model</i>			
Ours	95.9	9.43	3.30
Glow [30]	67.6	3.92	48.9
R-Flow [2]	70.3	3.60	46.4
IGEBM [6]	49.1	8.30	37.9
JEM [16]	92.9	8.76	38.4
<i>Explicit EBM</i>			
Diff Recovery [12]	N/A	8.30	9.58
VAEBM [57]	N/A	8.43	12.2
ImprovedCD [5]	N/A	7.85	25.1
CF-EBM [70]	N/A	-	16.7
EBMs-VAE [62]	N/A	6.65	36.2
CoopFlow [63]	N/A	-	15.8
CEM [54]	N/A	8.68	36.4
ATEBM [65]	N/A	9.10	13.2
HATEBM [20]	N/A	-	19.30
Adaptive CE [56]	N/A	-	65.01
CLEL [35]	N/A	-	8.61
<i>GANs</i>			
BigGAN [1]	N/A	9.22	14.7
SNGAN [40]	N/A	8.22	21.7
StyleGAN* [27]	N/A	8.99	9.90
<i>Score-Based Model</i>			
DDPM [21]	N/A	9.46	3.17
NCSN [50]	N/A	8.87	25.32
NCSN-v2 [51]	N/A	8.40	10.87
<i>Discriminative Model</i>			
Wide ResNet-28-12 [67]	95.6	N/A	N/A

Table 1: CIFAR-10 hybrid modeling results. ‘N/A’ means the corresponding result is not available. We report the result of *Wide ResNet-28-12*, which has a similar architecture, number of parameters and computational cost to our model.

the performance of the discriminative model Wide ResNet-28-12, which has a comparable architecture and number of parameters. Furthermore, the sampling quality of our model outperforms a majority of existing Explicit EBM, GAN models and Score-Based models, with IS of 9.43 and FID of 3.30. These results showcase the potential of EGC to enhance image classification performance and generative capability of EBM. On CIFAR-100, EGC model achieves comparable generative results with state-of-the-art GAN-based methods, while outperforming the hybrid model JEM [16] in terms of classification accuracy, as shown in Table 2.

We conducted additional experiments on the more challenging dataset, ImageNet. The results, presented in Table 3, provide evidence that our proposed EGC model performs well on this dataset, achieving IS of 189.5 FID of 6.77 and Top-1 Accuracy of 72.5%. The model was trained using random flip as data augmentation. We recognize

Method	Acc(%)	IS(\uparrow)	FID(\downarrow)
Ours	77.9	11.50	4.88
FQ-GAN [69]	N/A	7.15	9.74
LeCAM (BigGAN) [52]	N/A	-	11.2
StyleGAN2 + DA [28]	N/A	-	15.22
JEM [16]	72.2	-	-
Wide ResNet-28-12 [67]	79.5	N/A	N/A

Table 2: Results of EGC model on CIFAR-100 dataset.

Method	Acc(%)	IS(\uparrow)	FID(\downarrow)
EGC †	78.9	231.3	6.05
EGC ‡	72.5	189.5	6.77
EGC	70.4	79.9	17.5
ADM-Classifier [4]	64.3	N/A	N/A
HATEBM [20] (128×128)	N/A	-	29.37
IGEBM [6] (128×128)	N/A	28.6	43.7
IDDPM [42]	N/A	-	12.3
ADM [4]	N/A	100.98	10.94
LDM-VQ-8 [45]	N/A	201.56	7.77
VQGAN [8]	N/A	78.3	15.78

Table 3: Results of EGC model on ImageNet-1k 256×256 dataset using only random flip as data augmentation. † represents jointly training a conditional and an unconditional model, following [22]. ‡ represents incorporating the *RandResizeCrop* data augmentation. The sampling is guided by classifier gradient with a scale of 6 for EGC †‡ .

that incorporating stronger augmentation techniques could lead to even better results. We further jointly train a conditional and unconditional model, following [22], by randomly dropping the class label. The classification performance is evaluated with the unconditional model and the generation result is conducted with classifier guidance with a scale of 6. By incorporating the *RandResizeCrop* data augmentation, we achieve a significant increase of 78.9% in accuracy on the ImageNet dataset.

Unsupervised EGC model. EGC model can be trained in an unsupervised manner without knowing the label of image. By marginalizing out the variable y in Equation 14, the score function can be optimized by minimizing the Fisher divergence. As shown in Table 4, Unsupervised EGC model achieves FID of 7.75 on CelebA-HQ 256×256 and FID of 8.97 on LSUN Church 256×256 . These results demonstrate that Unsupervised EGC model outperforms the other state-of-the-art Energy-Based models. Although the score-based diffusion model, DDPM, exhibits slightly better performance, it is noteworthy that optimizing the gradient of the neural network for Unsupervised EGC model is a more challenging task. We believe that a network architecture specifically designed for optimizing the gradient will lead to the same or even better results than DDPM.

Ablation study. The results in Table 5 demonstrate

CelebA-HQ 256×256	FID(\downarrow)
Ours	7.75
ATEBM [65]	17.31
VAEBM [57]	20.38
CF-EBM [70] (128×128)	23.50
NVAE [53]	45.11
Glow [30]	68.93
ProgressiveGAN [26]	8.03

Table 4: Results of Unsupervised EGC models on the CelebA-HQ and LSUN Church datasets. Our model outperforms the other state-of-the-art Energy-Based Models in terms of FID.

EBM	Classifier	Guidance	Network	Acc(%)	FID(\downarrow)
✓			U-Net	N/A	5.36
✓	✓		U-Net	95.9	3.49
✓	✓	✓	U-Net	95.9	3.30
✓	✓	✓	ResNet	95.9	7.15

Table 5: The ablative results on CIFAR-10 dataset. EBM refers to Unsupervised EGC model model. Classifier represents EGC, and Guidance indicates the use of the gradient of the label for generating samples.

the effectiveness of the proposed EGC framework for image synthesis and classification. Unsupervised EGC model serves as a good baseline, achieving FID of 5.36. The EGC model achieves 95.9% accuracy on the test set and an additional 1.87 FID improvement, demonstrating the success of learning the joint probability. Moreover, we make use of class labels for conditional image synthesis. The classifier guidance $\nabla \log p_\theta(y|\mathbf{x}_t)$ guides the denoise process towards the class label y , resulting in a 0.19 FID improvement. To investigate the effect of neural network architecture, we train a model based on the standard feedforward ResNet commonly used for image classification. The comparison of the last and penultimate lines in Table 5 reveals that the U-Net architecture benefits from the short-cut connections specifically designed to propagate fine details from the inputs.

4.3. Application and Analysis

Interpolation. As shown in Figure 4, our model is capable of producing smooth interpolation between two generated samples. Following DDIM [49], we perform an interpolation between the initial white noise samples \mathbf{x}_T . We compare the generated samples from the interpolated noise with the samples obtained through interpolation in the generated samples \mathbf{x}_0 . The results demonstrate that our method exhibits superior semantic interpolation effects compared to direct interpolation of generated samples in latent space.

Image inpainting. A promising application of energy-based models is to use the learned prior model for filling masked regions of an image with new content. Following



Figure 4: Interpolation results between the leftmost and rightmost generated samples. \mathbf{x}_T denotes the interpolation the noise samples. \mathbf{x}_0 means the interpolation on the generated samples.



Figure 5: Inpainting results on CelebA-HQ dataset with shape of 256×256. The top row shows the masked images, while the bottom row displays the inpainted images.

[38], we obtain a sequence of masked noisy image at different timesteps and fill the masked pixels with the denoised sample given the previous iteration. The qualitative results on CelebA-HQ 256×256 are presented in Figure 5, demonstrating that our model is capable of realistic and semantically meaningful image inpainting.

Robustness. Adversarial attacks are a common threat to neural networks, especially when the model weights are accessible. A common white-box attack is FGSM, perturbing the inputs with the gradients enlarging the loss. We investigate the robustness of models trained on the CIFAR-10 dataset using FGSM [14] and PGD [39] attacks under the L_∞ constraint. The accuracy curve in Figure 7 demonstrates that the EGC model outperforms the standard Wide ResNet-28-12 classifier and JEM [16] in terms of adversarial robustness. The results suggest that leveraging the joint probability distribution learned by energy-based models can enhance the model’s robustness against adversarial attacks.

Visualize Energy. As detailed in Section 3.2, we adopt a direct optimization of the Fisher divergence instead of the probability $p_\theta(\mathbf{x}_t)$. Therefore, we are interested to see whether the neural network would effectively model the target Gaussian distribution. Given the difficulty in illustrating a high-dimensional Gaussian distribution, we present that the unnormalized probability density of noised sample \mathbf{x}_t in Figure 6a. Notably, the density exhibits a similar shape to the folded normal distribution, suggesting that the probability distribution learned by the neural network closely approximates the Gaussian distribution. In Figure 6b, we select two orthogonal noises to plot the two-dimensional

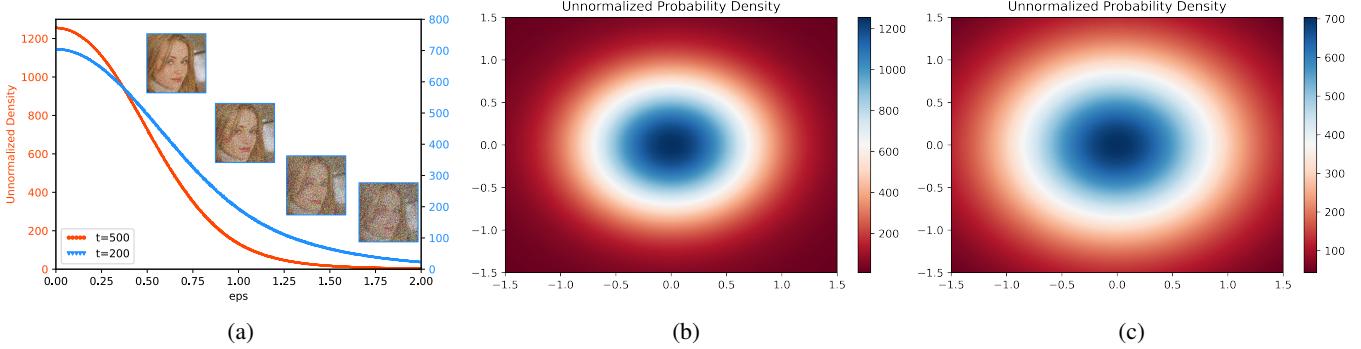


Figure 6: (a) The unnormalized probability density. The x-axis represents the noise level of \mathbf{x}_t , which is the mean of the abstract value of the noise at each timestep. The probability density exhibits a similar shape to the folded normal distribution. (b) The density function of the noised samples at $t = 500$. The noise is produced by linearly combining two orthogonal noises. The probability density exhibits a similar shape to the Gaussian distribution. (c) The density function of the noised samples at $t = 200$.

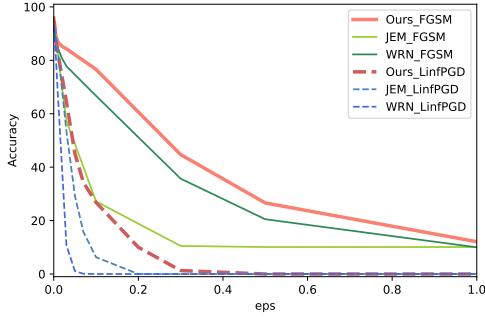


Figure 7: Robustness evaluation of EGC model on CIFAR-10 with PGD and FGSM adversarial attacks. Our proposed classifier exhibits considerable improvement in adversarial robustness compared to the baseline.

probability density function, which exhibits a similar shape to the Gaussian distribution.

Conditional sampling. As illustrated in Section 3.2, the classifier provides guidance to explicitly control the data we generate through conditioning information y . We feed the network with a fixed class label and random noise to check the qualitative results. As shown in Fig. 8, the diversity is promised by the random noise and the semantic consistency is guaranteed by the classification guidance. We further investigate the relationship between guidance scale with sample diversity. By increasing the guidance scale from 2 to 200, the diversity decreases significantly. An interesting observation is that the mode finally collapse to the class prototype. Meanwhile, the generated image with a high guidance scale are of high fidelity.

5. Conclusion

In this work, we introduce a novel energy-based model, EGC, to bridge the gap between discriminative and generative learning. We formulate the image classification and gener-

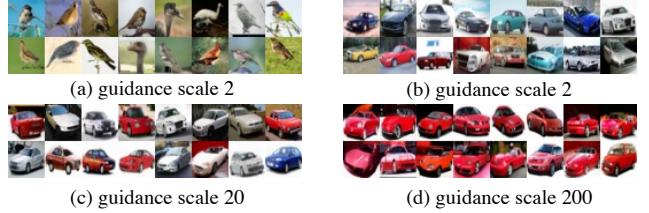


Figure 8: The conditional sampling results demonstrate that the random noise promises the diversity and the classification guidance guarantees the semantic consistency. As the guidance scale increase, the diversity significantly decreases and the mode collapse to the class prototype.

ation tasks from a probabilistic perspective, and find that the energy-based model is suitable for both tasks. To alleviate the difficult of optimizing the normalized probability density of energy-based model, we introduce diffusion process to populate the low data density regions for better estimated score, and score matching to circumvent optimizing the object loss with the normalizing constant $Z(\theta)$. In EGC, the forward pass models the joint distribution of the noisy image and label, while the backward pass of EGC calculates both conditional and unconditional scores to encourage the denoising process to generate samples consistent with the given labels. We achieve high-quality image synthesis and competitive image classification accuracy using a single neural network. We believe that EGC can serve as a new baseline for unifying discriminative and generative models.

Acknowledgement. This paper is partially supported by the National Key R&D Program of China No.2022ZD0161000, the General Research Fund of Hong Kong No.17200622 and Hong Kong Research Grants Council under Collaborative Research Fund (Project No. HKU C7004-22G).

References

- [1] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 6
- [2] Ricky TQ Chen, Jens Behrmann, David K Duvenaud, and Jörn-Henrik Jacobsen. Residual flows for invertible generative modeling. *Advances in Neural Information Processing Systems*, 32, 2019. 6
- [3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5
- [4] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 3, 5, 6
- [5] Yilun Du, Shuang Li, Joshua Tenenbaum, and Igor Mordatch. Improved contrastive divergence training of energy based models. *arXiv preprint arXiv:2012.01316*, 2020. 6
- [6] Yilun Du and Igor Mordatch. Implicit generation and generalization in energy-based models. *arXiv preprint arXiv:1903.08689*, 2019. 2, 6
- [7] Bradley Efron. Tweedie’s formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. 4
- [8] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 6
- [9] Chelsea Finn, Paul Christiano, Pieter Abbeel, and Sergey Levine. A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models. *arXiv preprint arXiv:1611.03852*, 2016. 2
- [10] Ruiqi Gao, Yang Lu, Junpei Zhou, Song-Chun Zhu, and Ying Nian Wu. Learning generative convnets via multi-grid modeling and sampling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9155–9164, 2018. 2
- [11] Ruiqi Gao, Erik Nijkamp, Diederik P Kingma, Zhen Xu, Andrew M Dai, and Ying Nian Wu. Flow contrastive estimation of energy-based models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7518–7528, 2020. 2
- [12] Ruiqi Gao, Yang Song, Ben Poole, Ying Nian Wu, and Diederik P Kingma. Learning energy-based models by diffusion recovery likelihood. *arXiv preprint arXiv:2012.08125*, 2020. 2, 6
- [13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2
- [14] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 7
- [15] Anirudh Goyal Alias Parth Goyal, Nan Rosemary Ke, Surya Ganguli, and Yoshua Bengio. Variational walkback: Learning a transition operator as a stochastic recurrent net. In *Advances in Neural Information Processing Systems*, pages 4392–4402, 2017. 2
- [16] Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. *arXiv preprint arXiv:1912.03263*, 2019. 2, 3, 6, 7
- [17] Qiushan Guo, Xinjiang Wang, Yichao Wu, Zhipeng Yu, Ding Liang, Xiaolin Hu, and Ping Luo. Online knowledge distillation via collaborative learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11020–11029, 2020. 3
- [18] Qiushan Guo, Zhipeng Yu, Yichao Wu, Ding Liang, Haoyu Qin, and Junjie Yan. Dynamic recursive neural network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5147–5156, 2019. 3
- [19] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5
- [20] Mitch Hill, Erik Nijkamp, Jonathan Mitchell, Bo Pang, and Song-Chun Zhu. Learning probabilistic models from generator latent spaces with hat ebm. *Advances in Neural Information Processing Systems*, 2022. 6
- [21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 3, 6, 7
- [22] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3, 6
- [23] Ziqi Huang, Kelvin CK Chan, Yuming Jiang, and Ziwei Liu. Collaborative diffusion for multi-modal face generation and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6080–6090, 2023. 3
- [24] Long Jin, Justin Lazarow, and Zhuowen Tu. Introspective classification with convolutional nets. In *Advances in Neural Information Processing Systems*, pages 823–833, 2017. 2, 3
- [25] Ge Kan, Jinhui Lü, Tian Wang, Baochang Zhang, Aichun Zhu, Lei Huang, Guodong Guo, and Hichem Snoussi. Bi-level doubly variational learning for energy-based latent variable models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18460–18469, 2022. 2
- [26] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 5, 7
- [27] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, 33:12104–12114, 2020. 6
- [28] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 6

- [29] Taesup Kim and Yoshua Bengio. Deep directed generative models with energy-based probability estimation. *arXiv preprint arXiv:1606.03439*, 2016. 2
- [30] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018. 6, 7
- [31] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5
- [32] Rithesh Kumar, Anirudh Goyal, Aaron Courville, and Yoshua Bengio. Maximum entropy generators for energy-based models. *arXiv preprint arXiv:1901.08508*, 2019. 2
- [33] Justin Lazarow, Long Jin, and Zhuowen Tu. Introspective neural networks for generative modeling. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2774–2783, 2017. 2, 3
- [34] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and Fujie Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006. 4
- [35] Hankook Lee, Jongheon Jeong, Sejun Park, and Jinwoo Shin. Guiding energy-based models via contrastive latent variables. 2023. 3, 6
- [36] Kwonjoon Lee, Weijian Xu, Fan Fan, and Zhuowen Tu. Wasserstein introspective neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3702–3711, 2018. 2, 3
- [37] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020. 2
- [38] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022. 7
- [39] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 7
- [40] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018. 2, 6
- [41] Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011. 2
- [42] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 6
- [43] Erik Nijkamp, Mitch Hill, Song-Chun Zhu, and Ying Nian Wu. On learning non-convergent short-run mcmc toward energy-based model. *arXiv preprint arXiv:1904.09770*, 2019. 2
- [44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5
- [45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 3, 5, 6
- [46] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 3
- [47] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 5
- [48] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 3
- [49] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 7
- [50] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019. 2, 3, 6
- [51] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020. 2, 6
- [52] Hung-Yu Tseng, Lu Jiang, Ce Liu, Ming-Hsuan Yang, and Weilong Yang. Regularizing generative adversarial networks under limited data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7921–7931, 2021. 6
- [53] Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in neural information processing systems*, 33:19667–19679, 2020. 7
- [54] Yifei Wang, Yisen Wang, Jiansheng Yang, and Zhouchen Lin. A unified contrastive energy-based model for understanding the generative ability of adversarial training. *arXiv preprint arXiv:2203.13455*, 2022. 6
- [55] Ze Wang, Jiang Wang, Zicheng Liu, and Qiang Qiu. Energy-inspired self-supervised pretraining for vision models. *arXiv preprint arXiv:2302.01384*, 2023. 2
- [56] Zhisheng Xiao and Tian Han. Adaptive multi-stage density ratio estimation for learning latent space energy-based model. *Advances in Neural Information Processing Systems*, 2022. 6
- [57] Zhisheng Xiao, Karsten Kreis, Jan Kautz, and Arash Vahdat. Vaebm: A symbiosis between variational autoencoders and energy-based models. *arXiv preprint arXiv:2010.00654*, 2020. 6, 7
- [58] Jianwen Xie, Yang Lu, Ruiqi Gao, and Ying Nian Wu. Cooperative learning of energy-based model and latent variable model via mcmc teaching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 3

- [59] Jianwen Xie, Yang Lu, Ruiqi Gao, Song-Chun Zhu, and Ying Nian Wu. Cooperative training of descriptor and generator networks. *IEEE transactions on pattern analysis and machine intelligence*, 42(1):27–45, 2018. 3
- [60] Jianwen Xie, Yang Lu, Song-Chun Zhu, and Yingnian Wu. A theory of generative convnet. In *International Conference on Machine Learning*, pages 2635–2644, 2016. 2, 3
- [61] Jianwen Xie, Zilong Zheng, Xiaolin Fang, Song-Chun Zhu, and Ying Nian Wu. Cooperative training of fast thinking initializer and slow thinking solver for conditional learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):3957–3973, 2021. 3
- [62] Jianwen Xie, Zilong Zheng, and Ping Li. Learning energy-based model with variational auto-encoder as amortized sampler. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10441–10451, 2021. 6
- [63] Jianwen Xie, Yaxuan Zhu, Jun Li, and Ping Li. A tale of two flows: Cooperative learning of langevin flow and normalizing flow toward energy-based model. *arXiv preprint arXiv:2205.06924*, 2022. 6
- [64] Zeyue Xue, Guanglu Song, Qiushan Guo, Boxiao Liu, Zhuofan Zong, Yu Liu, and Ping Luo. Raphael: Text-to-image generation via large mixture of diffusion paths. *arXiv preprint arXiv:2305.18295*, 2023. 3
- [65] Xuwang Yin, Shiying Li, and Gustavo K Rohde. Learning energy-based models with adversarial training. In *European Conference on Computer Vision*, pages 209–226. Springer, 2022. 2, 3, 6, 7
- [66] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 5
- [67] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 6
- [68] Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016. 2
- [69] Yang Zhao, Chunyuan Li, Ping Yu, Jianfeng Gao, and Changyou Chen. Feature quantization improves gan training. *arXiv preprint arXiv:2004.02088*, 2020. 6
- [70] Yang Zhao, Jianwen Xie, and Ping Li. Learning energy-based generative models via coarse-to-fine expanding and sampling. In *International Conference on Learning Representations*, 2020. 6, 7